*Biostatistics* (2009), **10**, 2, *pp.* 228–244
doi:10.1093/biostatistics/kxn029
Advance Access publication on August 28, 2008

# Biomarker evaluation and comparison using the controls as a reference population

YING HUANG*

*Fred Hutchinson Cancer Research Center, Public Health Sciences, 1100 Fairview Avenue North, M3-A410, Seattle, WA 98109, USA*
yhuang124@gmail.com

MARGARET SULLIVAN PEPE

*Fred Hutchinson Cancer Research Center, Public Health Sciences, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98109, USA*

SUMMARY

The classification accuracy of a continuous marker is typically evaluated with the receiver operating characteristic (ROC) curve. In this paper, we study an alternative conceptual framework, the "percentile value." In this framework, the controls only provide a reference distribution to standardize the marker. The analysis proceeds by analyzing the standardized marker in cases. The approach is shown to be equivalent to ROC analysis. Advantages are that it provides a framework familiar to a broad spectrum of biostatisticians and it opens up avenues for new statistical techniques in biomarker evaluation. We develop several new procedures based on this framework for comparing biomarkers and biomarker performance in different populations. We develop methods that adjust such comparisons for covariates. The methods are illustrated on data from 2 cancer biomarker studies.

*Keywords*: Biomarker; Classification; Covariate adjustment; Percentile value; ROC; Standardization.

## 1. INTRODUCTION

Molecular biotechnology may yield biomarkers for many purposes including early detection of disease, accurate sophisticated diagnosis, and monitoring of treatment effect. The development of biomarkers is a relatively recent area of research. Yet, the enormous investment of resources from public and private sectors testifies to the promise that this approach holds. The receiver operating characteristic (ROC) curve is typically used to describe the discriminatory capacity of a marker. However, most statisticians have limited familiarity with ROC methodology. Here, we use an alternative conceptual framework for marker evaluation that has very traditional statistical elements. We show that it has strong ties to ROC analysis, and importantly, we describe some new techniques afforded by this framework.

*To whom correspondence should be addressed.

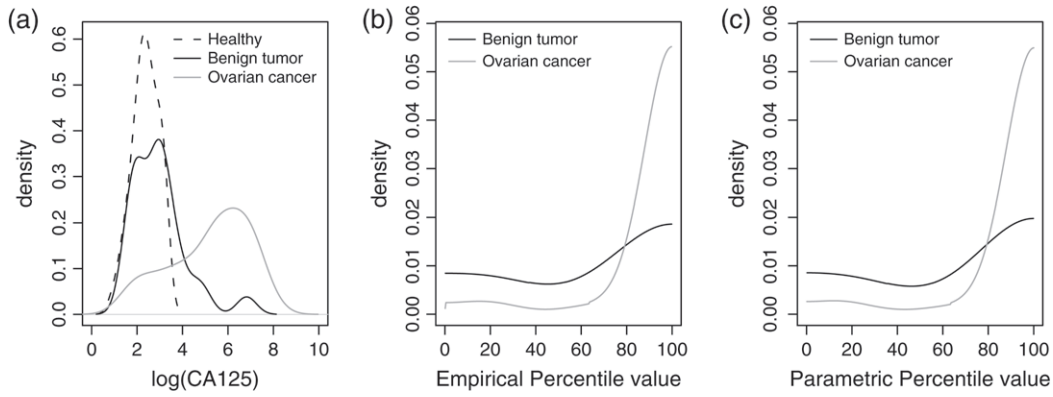


Fig. 1. Distributions of log(CA-125) in healthy women, women with benign tumors, and women with ovarian cancer (a); distributions of the estimated case percentile values when $F$ is estimated empirically (b) or parametrically (c).

Two specific problems are considered. The first is to determine if CA-125, a cancer antigen, discriminates women with benign ovarian tumors from healthy women as well as it discriminates women with clinically detected ovarian cancers from healthy women. Let $Y$ be the CA-125 measurement. Previously published data shown in Figure 1(a) are comprised of $\{Y_{\bar{D}i}, i = 1, \ldots, n_{\bar{D}}\}$ for controls, $\{Y_{1j}, j = 1, \ldots, n_1\}$ for cases with benign tumors, and $\{Y_{2j}, j = 1, \ldots, n_2\}$ for cases with ovarian cancer, where $n_{\bar{D}} = 41$, $n_1 = 24$, $n_2 = 66$, and $n_D = n_1 + n_2 = 90$ (McIntosh *and others*, 2004).

The second problem is to compare the discriminatory performances of 2 biomarkers, CA-19-9 and CA-125, for pancreatic cancer. For each of $n_D = 90$ cases with cancer and $n_{\bar{D}} = 51$ controls who did not have cancer but had pancreatitis (Wieand *and others*, 1989), the biomarkers denoted by $(Y_1, Y_2)$ are measured. The data are represented as $\{(Y_{1\bar{D}i}, Y_{2\bar{D}i}), i = 1, \ldots, n_{\bar{D}}, (Y_{1Dj}, Y_{2Dj}), j = 1, \ldots, n_D\}$.

We start by setting these 2 statistical problems in the new conceptual framework, without assuming any familiarity with ROC methodology. We develop several methods for inference, including a natural approach to covariate adjustment. Finally, we discuss how this framework relates to existing ROC methods and how it provides new methods for ROC analysis.

Proofs of theorems are given in Appendix B of the supplementary material (available at *Biostatistics* online, http://www.biostatistics.oxfordjournals.org).

## 2. Reference distribution standardization

The key idea is to use the biomarker distribution in controls as a reference distribution to standardize marker values. Let $F(Y)$ denote the cumulative distribution of the marker $Y$ in the control population. The standardized marker value, which we call its percentile value, is

$$\text{percentile value } = Q \equiv 100 \times F(Y). \tag{2.1}$$

This sort of standardization using a reference distribution is already commonplace in laboratory medicine and clinical medicine. In clinical medicine, for example, consider that weight and height of children are standardized relative to a healthy population of children of the same age and gender, so that reporting of percentile values is typical in practice (Frischancho, 1990).

Suppose without loss of generality that larger biomarker values are associated with disease (else we can use $-Y$ as the marker). An unusually large value of $Y$ has a percentile value close to 100. In laboratory medicine, a value of $Q$ above 95 or 99 might be flagged as outside the normal reference range. A good

biomarker would flag most cases as being outside the normal range. We propose that the distribution of case percentile values is a natural way to characterize the discriminatory performance of markers. On the one hand, with a useless marker the case and control distributions of $Y$ are the same so $Q$ has a uniform $(0, 100)$ distribution. On the other hand, an ideal marker will place all cases at $Q = 100$. The closer the case distribution of $Q$ is to that of the ideal, the better is the marker.

One could compare benign ovarian tumors and malignant cancers by their respective distributions of the standardized marker values. Substantially smaller values in benign tumor cases would indicate that discrimination is not as good for them as it is for malignant cancer cases. The standardization simplifies the problem by essentially reducing the number of groups from 3 to 2. In a sense, rather than evaluating if there is an interaction between disease status and disease type on $Y$, we need only do a simple 2-sample comparison of $Q$ between benign tumor cases and malignant cancer cases.

To compare 2 markers for discriminating a single set of cases from controls, each marker would be standardized with respect to its distribution in controls, yielding standardized values $Q_1$ and $Q_2$ for markers 1 and 2, respectively. If $Q_1$ tends to be larger than $Q_2$, marker 1 is the better marker because for cases it is more indicative of their disease than is marker 2. The standardization puts the 2 markers on a common scale where they can be compared using simple paired comparisons.

The approach of adopting the control distribution as a reference to standardize a biomarker has been taken in some biomarker studies (McIntosh *and others*, 2004) but has never been formalized as a valid statistical method. Moreover, since in practice only a finite sample of controls is available, formal statistical procedures need to acknowledge sampling variability in the reference distribution. We can estimate $F$ either empirically or parametrically with control data $\{Y_{\bar{D}i}, i = 1, \ldots, n_{\bar{D}}\}$. Write $\hat{F}$ for the estimator which in the setting of parametric estimation can also be written $F_{\hat{\theta}}$, where $\hat{\theta}$ is the estimated parameter for the model $F_\theta$. Even if marker values among cases are independent, their estimated standardized values, $\hat{Q}_j = 100 \times \hat{F}(Y_j)$, are not independent because of their common dependence on $\hat{F}$. This makes inference somewhat challenging.

## 3. COMPARING BENIGN TUMORS VERSUS OVARIAN CANCERS

### 3.1 *Comparing means*

*Unconditional test.* Let $Q_z(\hat{Q}_z)$ denote the percentile value (estimated) for the $z$th group of cases, with mean $E(Q_z)$, $z = 1, 2$. Let $\Delta = E(Q_1) - E(Q_2)$. The difference in sample means $\hat{\Delta} = \overline{\hat{Q}}_1 - \overline{\hat{Q}}_2$ can serve as the basis of a test statistic. Let $n_{\bar{D}}$ and $n_1$ and $n_2$ be the numbers of subjects in the control group and the first and second case groups, respectively.

THEOREM 3.1 Suppose marker observations are sampled independently and as $n_{\bar{D}} \to \infty, n_1/n_{\bar{D}} \to \lambda_1 \in (0, 1), n_2/n_{\bar{D}} \to \lambda_2 \in (0, 1)$, then $\sqrt{n_{\bar{D}}}(\hat{\Delta} - \Delta)$ converges to a mean 0 normal random variable with variance $\sigma^2$, where

$$(a) \quad \sigma^2 = \text{var}\{R_1(Y_{\bar{D}}) - R_2(Y_{\bar{D}})\} + \frac{\text{var}(Q_1)}{\lambda_1} + \frac{\text{var}(Q_2)}{\lambda_2}, \tag{3.1}$$

if $\hat{F}$ is the empirical cumulative distribution function (CDF), where $R_z(Y_{\bar{D}}) = P(Y_z < Y_{\bar{D}})$ denotes the percentile value of the marker $Y_{\bar{D}}$ from a control within the $z$th case distribution, and

$$(b) \quad \sigma^2 = \left(\frac{\partial \Delta}{\partial \theta}\right)^T \Sigma(\theta) \left(\frac{\partial \Delta}{\partial \theta}\right) + \frac{\text{var}(Q_1)}{\lambda_1} + \frac{\text{var}(Q_2)}{\lambda_2}, \tag{3.2}$$

if $F$ is modeled parametrically, where $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}(\hat{\theta} - \theta)$ and we assume that $\Delta$ is differentiable with respect to $\theta$.

Thus, the variability of $\hat{\Delta}$ comes from 2 sources, one due to sampling controls that form the reference population and the other due to sampling cases and calculating their percentile values given the reference distribution. In practice, we can estimate $\sigma^2$ using these formulas or the bootstrap method. If subjects are selected on the basis of their outcome status, resampling of subjects is done from the control and each case group separately. By calculating the variance of $\hat{\Delta} - \Delta$, we can construct a confidence interval (CI) for $\Delta$ and formally test for equality of $E(Q_1)$ and $E(Q_2)$.

In the ovarian cancer study (McIntosh *and others*, 2004), serum samples from 41 healthy women, 24 women with benign ovarian tumors, and 66 women with clinically detected ovarian cancer were assayed for CA-125. Figure 1(a) displays the distribution of log(CA-125) in the 3 groups. The difference between the ovarian cancer group and the healthy group is larger than the difference between the benign tumor group and the healthy group. We computed the percentile values of CA-125 in each of the case groups, using the empirical control distribution (Figure 1b) and under the assumption that log(CA-125) in controls follows a normal distribution after Box–Cox transformation (Figure 1c). Women with ovarian cancer appear to have larger percentile values of CA-125 compared to women with benign tumors.

Let $Q_1$ and $Q_2$ be percentile values for benign tumors and ovarian cancer groups, respectively. We calculated a 95% CI for $\Delta$. When $F$ is estimated empirically, $\overline{\hat{Q}}_1 = 63.31$, $\overline{\hat{Q}}_2 = 90.17$, $\hat{\Delta} = -26.86$, and the 95% CI for $\Delta$ is $(-42.77, -10.94)$ based on the asymptotic variance and $(-42.74, -10.97)$ based on the bootstrap variance. When $F$ is estimated parametrically, $\overline{\hat{Q}}_1 = 64.56$, $\overline{\hat{Q}}_2 = 90.03$, $\hat{\Delta} = -25.47$, and the 95% CI for $\Delta$ is $(-41.48, -9.46)$ based on the asymptotic variance and $(-41.39, -9.56)$ based on the bootstrap variance. Inferences based on the asymptotic and bootstrap variances agree fairly well here. The population mean percentile values are highly significantly different between the 2 case groups, regardless of how we model the marker distribution in controls (Table 1 in Appendix A of the supplementary material, available at *Biostatistics* online). The ability of CA-125 to identify ovarian cancer seems to be much better than its ability to detect benign tumors.

*Conditional test.* When our objective is hypothesis testing as opposed to estimation, we can consider testing for equality of mean percentile values conditional on the control sample. We use the term "conditional" inference here. The advantage of the conditional approach is that it maintains independence among the estimated percentile values, allowing standard 2-sample tests for independent samples to be applied when comparing case groups.

PROPOSITION 1 Using the notation $\stackrel{d}{=}$ for "equal distributions," under $H_0$: $Q_1 \stackrel{d}{=} Q_2$, if the support of the marker $Y$ in each case group is covered by its support in controls, then $Y_1 \stackrel{d}{=} Y_2$ and $\hat{Q}_1$ and $\hat{Q}_2$ have the same conditional distribution.

The implication of Proposition 1 is that if we reject the hypothesis that $\hat{Q}_1$ and $\hat{Q}_2$ have the same conditional distribution, we can reject the null hypothesis that $Q_1 \stackrel{d}{=} Q_2$. Earlier we used the unconditional test to compare the means of $Q_1$ and $Q_2$. In other words, we tested whether $E(\Delta) = 0$ where variability enters through both case and control samples. Here, we compare the means of $\hat{Q}_1$ and $\hat{Q}_2$ conditioning on the control sample. That is, we test whether $E(\hat{\Delta}|Y_i, i = 1, \ldots, n_{\bar{D}}) = 0$.

Observe that conditional on the control sample, the variance of $\hat{\Delta}$ is

$$\operatorname{var}\left( \frac{1}{n_1} \sum_{j=1}^{n_1} \hat{Q}_{1j} - \frac{1}{n_2} \sum_{i=1}^{n_2} \hat{Q}_{2j} \,\middle|\, Y_i, i = 1, \ldots, n_{\bar{D}} \right)$$

$$= \frac{\operatorname{var}(\hat{Q}_1|Y_i, i = 1, \ldots, n_{\bar{D}})}{n_1} + \frac{\operatorname{var}(\hat{Q}_2|Y_i, i = 1, \ldots, n_{\bar{D}})}{n_2}, \tag{3.3}$$

which can be consistently estimated by $\widehat{\text{var}}(\hat{Q}_1)/n_1+\widehat{\text{var}}(\hat{Q}_2)/n_2$, where $\widehat{\text{var}}$ denotes the sample variance. On the other hand, the unconditional variance of $\hat{\Delta}$ can be estimated by

$$\frac{\widehat{\text{var}}(\hat{Q}_1)}{n_1}+\frac{\widehat{\text{var}}(\hat{Q}_2)}{n_2}+\frac{\widehat{\text{var}}\{\hat{R}_1(Y_{\bar{D}})-\hat{R}_2(Y_{\bar{D}})\}}{n_{\bar{D}}} \geqslant \frac{\widehat{\text{var}}(\hat{Q}_1)}{n_1}+\frac{\widehat{\text{var}}(\hat{Q}_2)}{n_2}. \tag{3.4}$$

As a result, the conditional test comparing the means of $\hat{Q}_1$ and $\hat{Q}_2$ is always more powerful than the unconditional test. This is corroborated by the highly significant results in the top row of Table 1 in Appendix A of the supplementary material, available at *Biostatistics* online.

According to Proposition 1, $Q_1 \overset{d}{=} Q_2 \Leftrightarrow Y_1 \overset{d}{=} Y_2$. Therefore, an alternative way to test $H_0$: $Q_1 \overset{d}{=} Q_2$ is to compare the distributions of $Y_1$ and $Y_2$. Standard 2-sample tests for comparing 2 groups, such as the $t$-test, Wilcoxon rank sum test, or permutation test, all can be used for this purpose. Tests based on raw marker measurements and percentile values have the same type-I error under the null hypothesis but different powers under alternative hypotheses. In the example, the test comparing means of $Y_1$ and $Y_2$ is highly significant (Table 1 in Appendix A of the supplementary material, available at *Biostatistics* online), reaching the same conclusion as the test for equal means of $\hat{Q}_1$ and $\hat{Q}_2$, but this might not be true in other circumstances.

In summary, comparison of a marker's ability to differentiate 2 case groups from the same control group can be based on means of their percentile values $Q_1$ and $Q_2$. To construct a CI for $E(Q_1) - E(Q_2)$, we need to use unconditional inference that incorporates variability in controls as well as cases. On the other hand, simply to perform a hypothesis test for equality of the distributions of $Q_1$ and $Q_2$, the conditional methods should be used because of their enhanced power.

### 3.2 *Rank statistics*

Section 3.1 dealt with comparisons of mean percentile values. However, when distributions of percentile values do not belong to the same location-scale family (as shown in Figures 1b and c), alternatives to mean differences may be considered. For example, we can use rank-based statistics such as the Wilcoxon rank sum test, which is often used for comparing 2 groups of independent observations. For the problem at hand, we need to acknowledge the correlation among $\hat{Q}$s when applying the Wilcoxon rank sum test to them.

By analogy with methods in Section 3.1, we can apply the Wilcoxon rank sum test to $\hat{Q}_1$ and $\hat{Q}_2$ "unconditionally" or "conditional" on the control sample. In the former, the null hypothesis tested is $P(\hat{Q}_1 > \hat{Q}_2) = P(\hat{Q}_1 < \hat{Q}_2)$, which holds if $Q_1 \overset{d}{=} Q_2$ according to Proposition 1. In the latter, the null hypothesis tested is $P(\hat{Q}_1 > \hat{Q}_2|Y_i, i = 1, \ldots, n_{\bar{D}}) = P(\hat{Q}_1 < \hat{Q}_2|Y_i, i = 1, \ldots, n_{\bar{D}})$, which holds for all sets of control samples if $Q_1 \overset{d}{=} Q_2$. With the conditional testing, $\hat{Q}_1$ and $\hat{Q}_2$ are independent and the standard Wilcoxon rank sum test can be applied. For the unconditional test, the variance of the Wilcoxon rank sum test statistic can be estimated using the bootstrap.

In the ovarian cancer example, both the conditional and the unconditional Wilcoxon rank sum tests applied to $\hat{Q}$ suggest highly significant differences in the distributions of CA-125 percentile values between benign tumor cases and ovarian cancer cases (Table 1 in Appendix A of the supplementary material, available at *Biostatistics* online). Again, the conditional test is more powerful than the unconditional test since it does not involve variability in the control sample.

According to Proposition 1, we can also apply the Wilcoxon rank sum test to $Y_1$ and $Y_2$ to test the null hypothesis $Q_1 \overset{d}{=} Q_2$. Contrast the rank statistic based on $Y_1$ and $Y_2$ and that based on $\hat{Q}_1$ and $\hat{Q}_2$. If the transformation from $Y$ to $\hat{Q}$ does not change each observation's rank in the sample, then the rank-based statistic remains the same. This happens when $F$ is modeled as a strictly monotone increasing function but

does not necessarily happen when $F$ is estimated nonparametrically because ties may be created during the empirical CDF transformation. The increase in the number of ties will potentially affect the value of the test statistic and reduce its variance. For example, in the ovarian cancer data, the Wilcoxon rank sum test statistic applied to $Y_1$ and $Y_2$ has a value of $-524$ with a standard error 109.6, while the statistic applied to $\hat{Q}_1$ and $\hat{Q}_2$ has a value of $-437$ with a standard error 90.9 when $F$ is estimated empirically.

Note that if the nonparametric bootstrap is used for inference, the increase in ties during sampling with replacement can lead to underestimation of the variance. The severity of this problem depends on the sample size and the distribution of the percentile values. We found in limited simulation studies that for small sample size and good classification accuracy, applying the Wilcoxon rank sum test to $\hat{Q}$ with nonparametric bootstrap variance estimate led to anticonservative type-I error, especially when $F$ is estimated empirically. A solution is to use the smoothed bootstrap (Silverman, 1986; Silverman and Young, 1987). The idea is to simulate from smoothed distributions to avoid ties during resampling. There has been little systematic investigation about the choice of optimal bandwidth in this context. We explored several bandwidths in simulation studies and chose the bandwidth that covers around 40% of the total sample points in our data example. If variance estimation itself is not of interest, an alternative is to construct CIs based on percentiles of the nonparametric bootstrap distributions, an approach that turns out to be much less liberal than the Wald test based on nonparametric bootstrap variance estimates.

In summary, we can compare the discriminatory performance of a marker across different case groups using rank-based tests. We recommend (1) testing based on $\hat{Q}$ instead of $Y$ because the former is more relevant to differences in diagnostic accuracy and (2) using the conditional rather than the unconditional test because the former can be performed with standard statistical software and is more powerful, whereas the latter calls for smoothed bootstrap for variance estimation without a sound theoretic basis for bandwidth selection.

### 3.3 *Adjusting for covariates*

Suppose the biomarker distribution in controls varies with a covariate $X$ that can vary among cases, then the appropriate reference distribution should depend on $X$. We define the covariate-specific percentile value

$$Q_X = 100 \times F(Y|X), \tag{3.5}$$

where $F(Y|X)$ is the CDF of the marker in the control population with covariate value $X$. In clinical medicine, for anthropometric measurements it is standard practice to calculate covariate-specific percentile values. For example, the percentiles of height for children are age and gender specific because these factors affect height in normal healthy children. Berres *and others* (2008) described methods to estimate covariate-specific diagnostic scores.

To compare women with benign tumors and women with ovarian cancer, we can evaluate covariate-specific percentile values for each case group and compare them using 2-sample test statistics. Is covariate adjustment important? The answer is "potentially yes." Suppose, for example, that $X$ is age and that in controls older age is associated with larger values of the biomarkers. If women with ovarian cancer tend to be older than women with benign tumors, one would observe a difference in discriminatory performance that is simply due to age. Using age-adjusted biomarker percentiles is a simple way to eliminate such confounding.

If $X$ is discrete and there are relatively large numbers of controls per $X$ category, a nonparametric approach to estimating $F(Y|X)$ can be taken. Otherwise a parametric model is employed. For $z = 1, 2$, let $Q_{zX}(\hat{Q}_{zX})$ be the (estimated) covariate-specific percentile value for an observation in case group $z$. Let $\Delta = E(Q_{1X}) - E(Q_{2X})$ and $\hat{\Delta} = \overline{\hat{Q}}_{1X} - \overline{\hat{Q}}_{2X}$. When covariate $X$ is discrete with $K$ categories, let $n_{\bar{D}k}$ and $n_{zk}$ be the number of controls and the number of $z$th type of cases in the $k$th covariate category, $k = 1, \ldots, K$.

THEOREM 3.2 Suppose $n_{\bar{D}} \to \infty, n_1/n_{\bar{D}} \to \lambda_1 \in (0,1)$, and $n_2/n_{\bar{D}} \to \lambda_2 \in (0,1)$. When $X$ is discrete, suppose $n_{\bar{D}k}/n_{\bar{D}} \to p_{\bar{D}k} \in (0,1)$, $n_{1k}/n_1 \to p_{1k} \in (0,1)$, and $n_{2k}/n_2 \to p_{2k} \in (0,1)$. Then $\sqrt{n_{\bar{D}}}(\hat{\Delta} - \Delta)$ converges to a mean 0 normal random variable with variance $\sigma^2$, where

$$(a) \quad \sigma^2 = \sum_k \left[ \frac{\text{var}\{R_1^k(Y_{\bar{D}}^k)\}}{p_{\bar{D}k}/p_{1k}^2} + \frac{\text{var}\{R_2^k(Y_{\bar{D}}^k)\}}{p_{\bar{D}k}/p_{2k}^2} \right] + \frac{\text{var}(Q_{1X})}{\lambda_1} + \frac{\text{var}(Q_{2X})}{\lambda_2}, \tag{3.6}$$

if $F(Y|X)$ is modeled with the empirical CDF within the $k$th covariate category, where $R_z^k(Y_{\bar{D}}^k) = P(Y_z^k < Y_{\bar{D}}^k)$ and the $k$ superscript indicates cases and controls in covariate category $k$, and

$$(b) \quad \sigma^2 = \left(\frac{\partial \Delta}{\partial \theta}\right)^T \Sigma(\theta) \left(\frac{\partial \Delta}{\partial \theta}\right) + \frac{\text{var}(Q_{1X})}{\lambda_1} + \frac{\text{var}(Q_{2X})}{\lambda_2}, \tag{3.7}$$

if $F(Y|X)$ is modeled parametrically, where $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}(\hat{\theta} - \theta)$ and we assume that $\Delta$ is differentiable with respect to $\theta$ and that $\mathcal{F} = \{F_\theta(y|x): \theta \in \Theta\}$ is a Donsker (1952) class.

To illustrate, we simulated a continuous covariate $X$ for the ovarian cancer data. $X$ is generated to be positively associated with both CA-125 and disease status, $X \sim N(\mu, \sigma)$ where $\mu = 10 \times \log\{5 \times I(\text{benign tumors}) \times I(\log(\text{CA-125}) > 2.2) + 0.8 \times I(\text{ovarian cancer}) + 1.5 \times \log(\text{CA-125})\}$ and $\sigma = 4$. Figure 2 shows the distribution of log(CA-125) ignoring covariate $X$ and when $X$ is equal to its first, second, and third quartiles in the whole sample. Observe that the distribution of log(CA-125) in controls varies with $X$. Moreover, the separations between controls and case groups differ with $X$.

We calculated the covariate-specific percentile values assuming normality of log(CA-125) in controls conditional on $X$. The mean is modeled as a cubic B-spline in $X$, with pre-chosen knots at the first 3 quartiles in the control sample. Figure 3 plots the distributions of the marginal and covariate-specific percentile values of CA-125 for women in the 2 case groups. It appears that adjusting for the covariate $X$ reduces the separation between women with benign tumors and healthy women, while the separation between women with ovarian cancer and healthy women is unchanged. Indeed, the covariate-specific percentile values have an approximately uniform (0, 100) distribution for women with benign tumors indicating that their distribution is the same as that for controls. Therefore, covariate adjustment appears to be desirable in this setting. After covariate adjustment, CA-125 picks up fewer benign tumor cases while maintaining its ability to identify ovarian cancer cases.

We now formally compare the 2 groups of cases with regard to their covariate-specific percentile values. All the unconditional tests described in Sections 3.1 and 3.2 can be applied. All tests suggest that CA-125 has significantly better discriminatory performance for identifying ovarian cancer compared to benign tumors (Table 2 in Appendix A of the supplementary material, available at *Biostatistics* online). In terms of estimation, we find that, as expected for benign tumors, $\overline{\hat{Q}}_{1X}$ is close to the uninformative marker value of 50 ($\overline{\hat{Q}}_{1X} = 50.13$). In the ovarian cancer group, $\overline{\hat{Q}}_{2X} = 88.10$ which is similar to the mean unadjusted percentile values ($\overline{\hat{Q}}_2 = 90.17$). The difference in the covariate-adjusted means is $\hat{\Delta} = -37.96$, with 95% CI $(-57.76, -18.16)$ based on the asymptotic variance and $(-58.79, -17.13)$ based on the bootstrap variance.

In summary, when the marker distribution in controls varies with a covariate that can vary among cases, covariate-specific percentile values can be calculated to eliminate potential confounding. The 2 groups of cases can then be compared using mean or rank-based statistics. This provides a covariate-adjusted comparison of the discriminatory capacity of the marker. See Janes and Pepe (2008a,c) for a broad discussion of covariate adjustment.

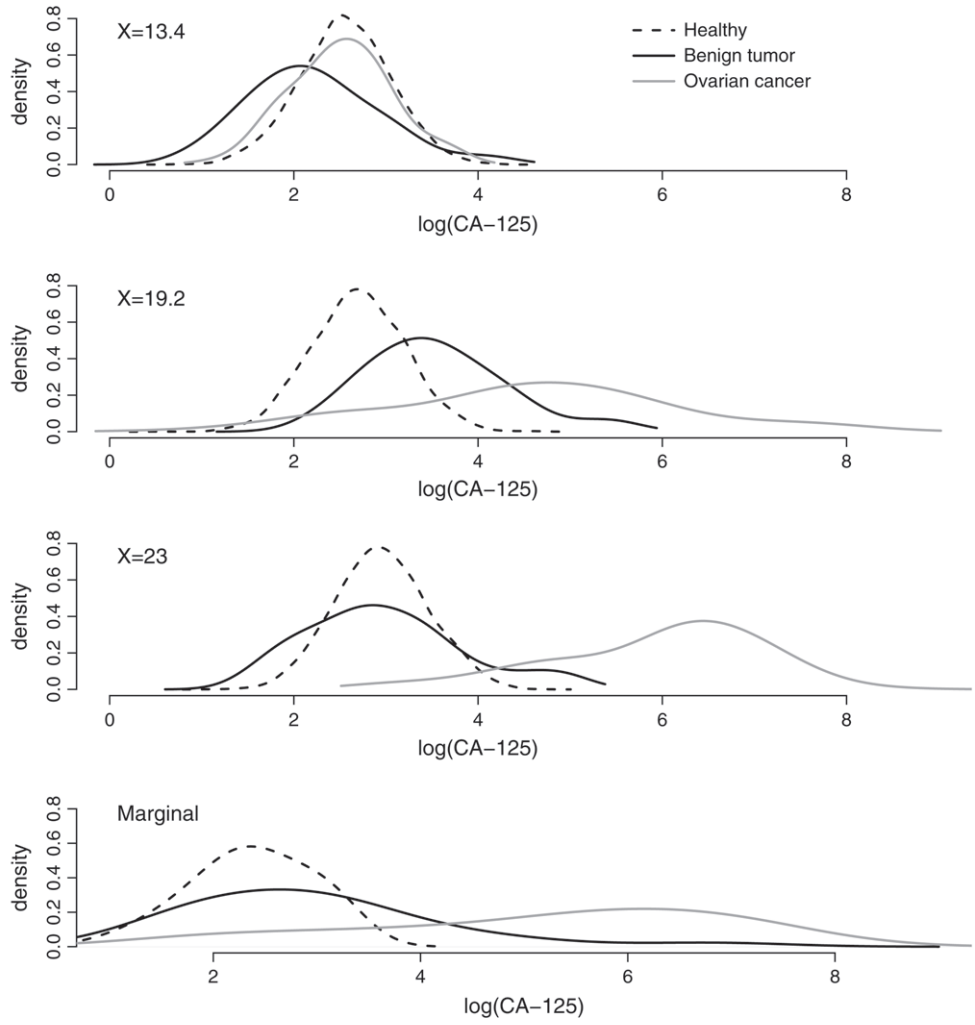


Fig. 2. Marginal and covariate-specific distributions of log(CA-125) in healthy women, women with benign ovarian tumors, and women with ovarian cancer.

## 4. Comparing markers

Next, consider the comparison of 2 markers with respect to their diagnostic accuracies. Two markers are measured on each of $n_D$ cases and $n_{\bar{D}}$ controls. Let $F_z, z = 1, 2$, be the distribution function for the $z$th marker in controls, and let $Q_z(\hat{Q}_z)$ denote the corresponding (estimated) case percentile value. Observe that each marker is standardized with respect to its own control reference distribution. Even though the raw marker values may be in different units, the transformation to percentile values put them on the same scale.

### 4.1 *Using means*

For each case, one can compare $Q_1$ and $Q_2$. If $Q_1$ tends to be larger than $Q_2$, then marker 1 is the better marker. Formally, let $\Delta = E(Q_1) - E(Q_2)$. The difference in sample means can serve as the basis of a test statistic $\hat{\Delta} = \overline{\hat{Q}}_1 - \overline{\hat{Q}}_2$.

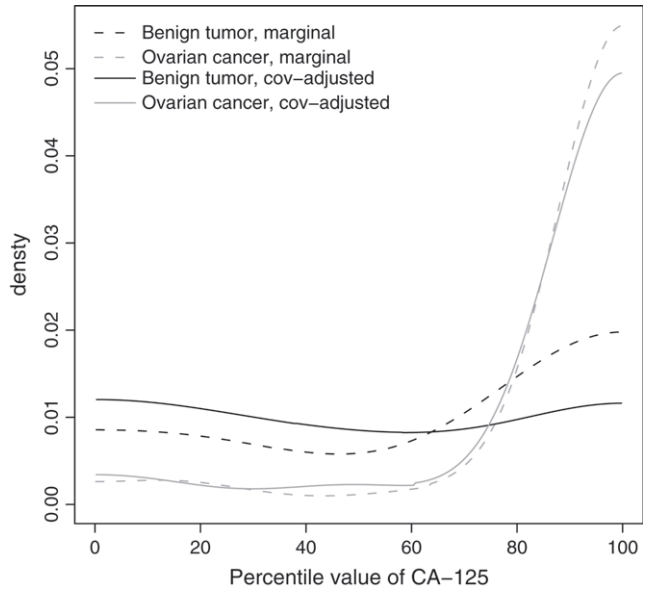


Fig. 3. Marginal and covariate-adjusted distributions of estimated percentile values of CA-125 for women with benign ovarian tumors and women with ovarian cancer.

In this 2-marker setting, correlation between the estimated percentile values comes from 2 sources: one due to subject-specific effects and the other due to estimation of the reference distributions. We need to acknowledge this correlation in making inference.

THEOREM 4.1 Suppose $n_D/n_{\bar{D}} \to \lambda$ as $n_{\bar{D}} \to \infty$, then $\sqrt{n_{\bar{D}}}(\hat{\Delta} - \Delta)$ converges to a mean 0 normal random variable with variance $\sigma^2$, where

$$(a) \quad \sigma^2 = \text{var}\{R_1(Y_{1\bar{D}}) - R_2(Y_{2\bar{D}})\} + \frac{\text{var}(Q_1 - Q_2)}{\lambda}, \tag{4.1}$$

if $F_z$ is estimated with the empirical CDF, where $Y_{z\bar{D}}$ and $Y_{zD}$ are measurements of the $z$th marker for a control and a case, respectively, and $R_z(Y_{z\bar{D}}) = P(Y_{zD} < Y_{z\bar{D}})$ is the percentile value of the $z$th marker from a control in the corresponding case distribution (Delong *and others*, 1988), and

$$(b) \quad \sigma^2 = \left(\frac{\partial \Delta}{\partial \theta}\right)^T \Sigma(\theta) \left(\frac{\partial \Delta}{\partial \theta}\right) + \frac{\text{var}(Q_1 - Q_2)}{\lambda}, \tag{4.2}$$

if $F_z$ is modeled parametrically with parameter $\theta_z$, where $\theta = (\theta_1, \theta_2)$, $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}(\hat{\theta} - \theta)$, and we assume that $\Delta$ is differentiable with respect to $\theta$.

In practice, $\sigma^2$ can be estimated based on these formulas or by bootstrap resampling.

Observe that, for this 2-marker problem, conditional inference is no longer applicable. Even if the distributions of $Q_1$ and $Q_2$ are the same, the distributions of $\hat{Q}_1$ and $\hat{Q}_2$ conditional on the particular control sample will not necessarily be equal. Therefore, testing the null hypothesis that $\hat{Q}_1|\{Y_{\bar{D}i}, i = 1, \ldots, n_{\bar{D}}\} \stackrel{d}{=} \hat{Q}_2|\{Y_{\bar{D}i}, i = 1, \ldots, n_{\bar{D}}\}$ is not equivalent to testing the null hypothesis that $Q_1 \stackrel{d}{=} Q_2$.

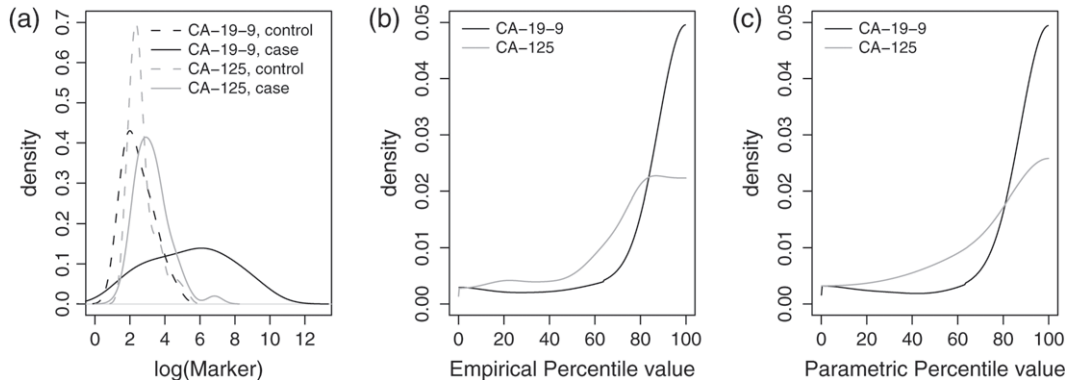


Fig. 4. Distributions of log(CA-19-9) and log(CA-125) in controls and cases (a); distributions of the estimated case percentile values when control distributions are estimated empirically (b) or parametrically (c).

The data set we use for illustration here is from the pancreatic cancer serum biomarker study (Wieand *and others*, 1989), which includes 90 cases and 51 controls. Serum samples from each patient were assayed for CA-19-9, a carbohydrate antigen, and CA-125, a cancer antigen.

Figure 4(a) shows the probability distributions of the markers. Also displayed are the distributions of the estimated case percentile values for each marker, with $F_z$ estimated empirically in Figure 4(b), and under the assumption that $Y$ is normally distributed after Box–Cox transformation in Figure 4(c). Clearly, the distribution of the percentile values for CA-19-9 is shifted to the right compared with CA-125, indicating that it is a better biomarker.

Next, consider the mean percentile values. When $F_z$ is estimated empirically, $\overline{\hat{Q}}_1 = 86.23$ for CA-19-9, $\overline{\hat{Q}}_2 = 70.70$ for CA-125, and $\hat{\Delta} = 15.53$. The corresponding 95% CI for $\Delta$ is $(4.34, 26.73)$ using the asymptotic variance and similarly $(4.37, 26.70)$ using the bootstrap variance. When $F_z$ is estimated parametrically, results are similar: $\overline{\hat{Q}}_1 = 86.07$, $\overline{\hat{Q}}_2 = 71.09$, and $\hat{\Delta} = 14.97$. The corresponding 95% CI for $\Delta$ is $(3.80, 26.15)$ using the asymptotic variance and $(3.57, 26.38)$ using the bootstrap variance. CA-19-9 performs significantly better than CA-125 for diagnosing pancreatic cancer (see also Table 2 in Appendix A of the supplementary material, available at *Biostatistics* online, for $p$-values).

In summary, to compare the diagnostic accuracy of 2 markers, we can use the controls to standardize the marker values in cases and compare the corresponding means. If $n_{\bar{D}} = \infty$, this is essentially a paired $t$-test. If $n_{\bar{D}} < \infty$, the paired $t$-test needs to be modified to accommodate the additional variability in the estimated control marker distributions.

### 4.2 *Using rank statistics*

Rank-based tests provide another avenue to compare the distributions of percentile values. Due to their complicated correlation structure, standard variance formulas for rank-based test statistics no longer apply. The bootstrap method is used instead. Moreover, as discussed earlier, conditional tests are not applicable here. So only unconditional tests are considered.

PROPOSITION 2 Under $H_0$: $Q_1 \overset{d}{=} Q_2$, we have $\hat{Q}_1 \overset{d}{=} \hat{Q}_2$ when $F_z$ is estimated empirically.

PROPOSITION 3 Let $U_j = \hat{Q}_{1j} - \hat{Q}_{2j}$, $j = 1, \ldots, n_D$. Let $T$ and $S$ be the Wilcoxon signed rank test statistic and the Sign test statistic, respectively. Under $H_0$: $Q_1 \overset{d}{=} Q_2$, we have $E(T) = (n_D + 1)/4$ and $E(S) = 1/2$ when $F_z$ is estimated empirically.

PROPOSITION 4 Let $r_k$ be the rank of $\tilde{Q}_k$, where

$$\{\tilde{Q}_k, k = 1, \ldots, 2n_D\} = \{\hat{Q}_{1j}, j = 1, \ldots, n_D, \hat{Q}_{2j}, j = 1, \ldots, n_D\}.$$

Let $W = \sum_{k=1}^{n_D} r_k$ be the Wilcoxon rank sum test statistic. Then under $H_0$: $Q_1 \stackrel{d}{=} Q_2$, $E(W) = n_D(2n_D + 1)/12$ when $F_z$ is estimated empirically.

We expect the results in Propositions 2–4 to hold asymptotically when $F_z$ is estimated parametrically. In other words, under $H_0$: $Q_1 \stackrel{d}{=} Q_2$, the expectations of these rank-based test statistics applied to $\hat{Q}_1$ and $\hat{Q}_2$ are the same as that in the standard 2-sample setting (for $W$) and the paired-data setting (for $T$ and $S$). Therefore, to test for equal discriminatory performance of 2 markers, we can apply the rank-based test statistics to $\hat{Q}_1$ and $\hat{Q}_2$, bootstrapping the variance. Here, we face the same concerns about underestimation of the variance as in Section 3.2. Using the smoothed bootstrap for variance estimation or constructing CIs based on nonparametric bootstrap distributions seems to be a solution. Asymptotic distribution theory appears to be very challenging. Using a smoothed bootstrap with a bandwidth covering approximately 40% sample points, all rank-based tests suggest a highly significant difference between the 2 markers (Table 2 in Appendix A of the supplementary material, available at *Biostatistics* online).

### 4.3 *Adjusting for covariates*

We argued earlier that adjusting for covariates may be important when comparing 2 case groups. This is also potentially important when comparing 2 biomarkers. Suppose, for example, that biomarker values in the control group vary with study site in a multicenter study. Such might occur if collection or processing procedures differed across sites. If the site-specific control populations are pooled to form a reference set, the distribution of the case percentiles may be more diffuse than if the site-specific controls are used as the reference group (see the right side of Figure 5 for an example). Even if the case–control ratio is the same across study sites, biomarker performance can appear to be worse than it is by using a pooled reference set (Janes and Pepe, 2008b). Markers may differ with regard to this phenomenon. For example, processing techniques that vary across sites may affect one marker but not another. Differential covariate effects on reference distributions of biomarkers therefore can bias the comparison of markers unless proper adjustment is undertaken. The use of covariate-specific percentile values is a means to adjust for covariates and avoid this bias. Note that pertinent covariates may be different for different markers.

For $z = 1, 2$, let $Q_{zX}(\hat{Q}_{zX})$ be the (estimated) covariate-specific percentile value for the $z$th marker, $\Delta = E(Q_{1X}) - E(Q_{2X})$ and $\hat{\Delta} = \overline{\hat{Q}}_{1X} - \overline{\hat{Q}}_{2X}$. When $X$ is discrete with $K$ categories, let $n_{\bar{D}k}$ and $n_{Dk}$ be the numbers of controls and cases in the $k$th covariate category, $k = 1, \ldots, K$. Again covariate adjustment is only relevant when the covariate is defined for both cases and controls.

THEOREM 4.2 Suppose as $n_{\bar{D}} \to \infty$, $n_D/n_{\bar{D}} \to \lambda \in (0, 1)$, and for a discrete covariate, $n_{\bar{D}k}/n_{\bar{D}} \to p_{\bar{D}k} \in (0, 1)$, $n_{Dk}/n_D \to p_{Dk} \in (0, 1)$, then $\sqrt{n_{\bar{D}}}(\hat{\Delta} - \Delta)$ converges to a mean 0 normal random variable with variance $\sigma^2$, where

$$(a) \quad \sigma^2 = \sum_k \frac{\text{var}\{R_1^k(Y_{1\bar{D}}^k) - R_2^k(Y_{2\bar{D}}^k)\}}{p_{\bar{D}k}/p_{Dk}^2} + \frac{\text{var}(Q_{1X} - Q_{2X})}{\lambda}, \tag{4.3}$$

if the covariate-specific reference distribution $F(Y|X)$ is estimated empirically within each covariate category, where $R_z^k(Y_{z\bar{D}}^k) = P(Y_{zD}^k < Y_{z\bar{D}}^k)$ is the percentile value for a control using its covariate-specific

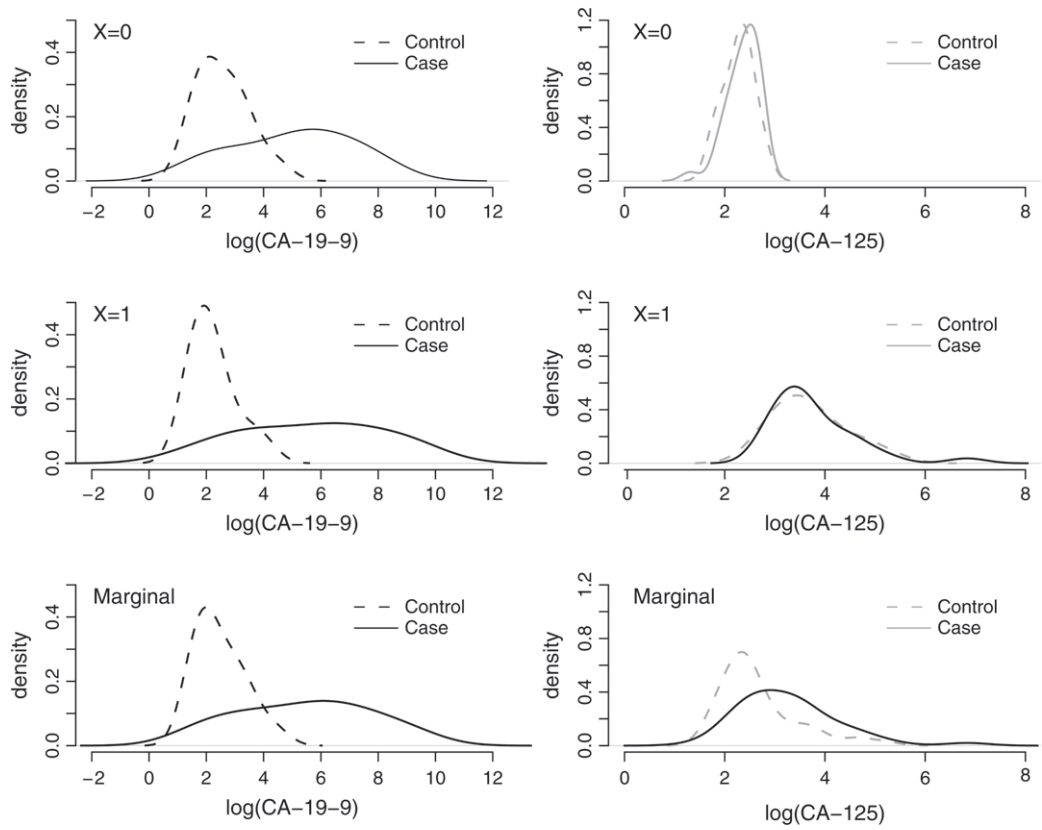


Fig. 5. Marginal and covariate-specific distributions of log(CA-19-9) and log(CA-125) in controls and cases.

case distribution as the reference for the $z$th marker in the $k$th covariate category, and

$$(b) \quad \sigma^2 = \left(\frac{\partial \Delta}{\partial \theta}\right)^T \Sigma(\theta) \left(\frac{\partial \Delta}{\partial \theta}\right) + \frac{\text{var}(Q_{1X} - Q_{2X})}{\lambda}, \tag{4.4}$$

if $F(Y|X)$ is modeled parametrically for marker $z$ with parameter estimate $\theta_z$, where $\theta = (\theta_1, \theta_2)$ and $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}(\hat{\theta} - \theta)$. We assume that $\Delta$ is differentiable with respect to $\theta$ and that $\mathcal{F} = \{F_\theta(y|x) : \theta \in \Theta\}$ is a Donsker class.

As shown in the supplementary material, available at *Biostatistics* online, Theorem 4 extends to the setting when different covariates are used to adjust different markers.

To illustrate, we simulate a discrete covariate $X$ for the pancreatic cancer data. We set $X$ to 1 for those with CA-125 above its median, and 0 otherwise. In total, 14 out of 51 (27.4%) controls and 57 out of 90 (63.3%) cases have $X = 1$. Figure 5 shows the probability distributions of log(CA-19-9) and log(CA-125) conditional on $X$.

For CA-19-9, the value of $X$ does not have a dramatic influence on the reference control distribution, suggesting that covariate adjustment is not warranted. On the other hand, for CA-125, since the marker is positively associated with $X$ and a higher percentage of cases have $X = 1$ compared with controls, the distribution for cases shifts to the right compared to the distribution for controls when data are pooled over

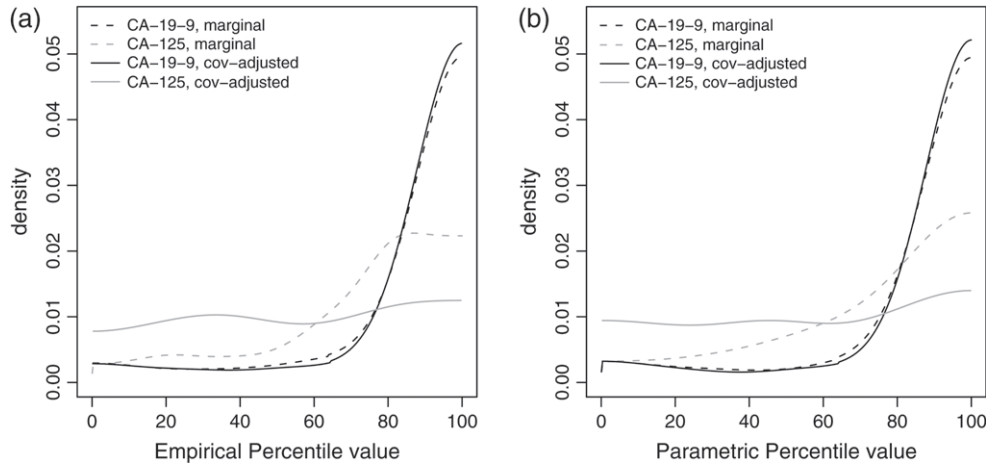


Fig. 6. Marginal and covariate-adjusted distributions of the estimated case percentile values of CA-19-9 and CA-125.

$X$, even if there is not much difference between them conditional on $X$. In other words, $X$ is a confounder for CA-125 but not for CA-19-9. Distributions of the covariate-specific percentile values for CA-19-9 ($\hat{Q}_{1X}$) and CA-125 ($\hat{Q}_{2X}$) in cases are shown in Figure 6. For CA-19-9, covariate adjustment does not affect the distribution of the case percentile values, whereas for CA-125, covariate adjustment removes the confounding effect of $X$ and suggests performance that is poorer than its marginal performance.

With $F(Y|X)$ estimated empirically, we have $\overline{\hat{Q}}_{1X} = 87.25$ for CA-19-9, $\overline{\hat{Q}}_{2X} = 53.85$ for CA-125, and $\hat{\Delta} = 33.40$. The corresponding 95% CI for $\Delta$ is $(20.04, 46.76)$ using the asymptotic variance and $(20.83, 45.97)$ using the bootstrap variance. With $F(Y|X)$ estimated parametrically under the assumption that $Y$ is normally distributed after Box–Cox transformation within each covariate category, we find $\overline{\hat{Q}}_{1X} = 87.09$, $\overline{\hat{Q}}_{2X} = 54.20$, and $\hat{\Delta} = 32.89$. The corresponding 95% CIs for $\Delta$ are $(18.97, 46.81)$ and $(20.38, 45.40)$ using the asymptotic and bootstrap variances, respectively. See Table 2 in Appendix A of the supplementary material, available at *Biostatistics* online, for $p$-values based on mean and rank statistics. CA-19-9 appears to be a much better marker than CA-125 for identifying pancreatic cancer, especially after adjusting for the covariate.

## 5. RELATIONSHIPS WITH ROC ANALYSIS

Our approach to evaluating the capacity of a marker to distinguish cases from a reference set of controls is to use the control marker distribution to standardize marker values for cases. If these percentile values tend to be high for many cases, the marker's discriminatory capacity is good. We noted earlier that the approach is intuitive and is used in some applications (McIntosh *and others*, 2004). Interestingly, it is equivalent to ROC analysis, which plays a central role in biomarker evaluation (Baker, 2003; Pepe, 2003). The equivalence has been noted previously (Pepe and Cai, 2004; Pepe and Longton, 2005). In particular, since the ROC curve, a plot of true-positive rate (TPR) $= P(Y > c|D = 1)$ versus false-positive rate (FPR) $= P(Y > c|D = 0)$, can be written as

$$\mathrm{ROC}(t) = P\{Y > S^{-1}(t)|D = 1\} = P\{S(Y) < t|D = 1\}, \quad t \in (0, 1), \tag{5.1}$$

where $S = 1 - F$, we see that the ROC curve is the CDF of $1 - F(Y)$ in cases. Thus, comparing case distributions of biomarker percentile values, $Q = 100 \times F(Y)$, is entirely equivalent to comparing ROC

curves. The representation of the distribution of $Q$ in terms of the ROC curve provides further justification for using case percentile values as the unit of analysis in evaluating and comparing markers. Empirical ROC curves for the ovarian and pancreatic cancer data sets are shown in Figure 1 in Appendix A of the supplementary material, available at *Biostatistics* online.

Some of the procedures presented in Sections 3 and 4 are alternative representations of existing procedures for comparing ROC curves while some are new procedures. Using the fact that the mean of a random variable is equal to the area under its survival function, we see that the average of case percentile values can be represented in terms of the area under the ROC curve (AUC) (Bamber, 1975),

$$\text{AUC} = E(Q)/100. \tag{5.2}$$

Thus, comparisons based on mean percentile values are equivalent to comparisons of AUCs, the classical approach to comparing ROC curves.

Hanley and Hajian-Tilaki (1997) represented the empirical AUC as the sample mean of case percentile values with $F$ estimated empirically. The asymptotic results in Theorems 1(a) and 2(a) are results for empirical AUC differences that have been previously reported (Sukhatme and Beam, 1994; Delong *and others*, 1988). However, their semiparametric counterparts in Theorems 1(b) and 2(b) have not. Li *and others* (1996) studied semiparametric estimation of the ROC curve when the case distribution is modeled parametrically and the control distribution is modeled empirically. We did the reverse in this paper using a flexible smooth form for the reference control distribution. The Box–Cox family has precedent in modeling reference distributions for anthropometric measures (Cole, 1990). Returning to the asymptotic results in Theorems 1(a) and 2(a), in contrast to Sukhatme and Beam (1994) and similar to Hanley and Hajian-Tilaki (1997), we reparameterized the variances in terms of percentile values in this report, which we feel is a more intuitive way to understand the components of the variance.

A problem with comparing the diagnostic accuracy of 2 tests using AUC is lack of power to detect differences in ROC curves when they have the same area under the curve. As pointed out by Swets (1986), ROC curves are typically asymmetric, and 2 ROC curves with different asymmetries might cross each other but have the same AUC. Venkatraman and Begg (1996) developed a permutation test procedure to compare 2 ROC curves with paired data. Extension of the permutation test to the setting of continuous unpaired data was also proposed (Venkatraman, 2000). Extension to comparisons among more than 2 tests, however, might be computationally intensive.

The rank statistics described in Sections 3.2 and 4.2 provide an alternative solution to distinguishing between curves with the same AUCs. They have power to reject $H_0$: $Q_1 \stackrel{d}{=} Q_2$ when $E(Q_1) = E(Q_2)$ but $P(Q_1 > Q_2) \neq P(Q_1 < Q_2)$. These can be interpreted as new ROC analysis techniques. Yet, their interpretation as rank statistics to compare distributions of standardized biomarkers in cases is equally valid and may be preferred by some. The generalization to comparing distributions of multiple standardized biomarkers is also tenable (Cuzick, 1985; Kruskal and Wallis, 1952).

Nakas *and others* (2003) proposed comparing markers using functions of case percentile values. Their statistic is a nonstandard ROC summary index, namely, the 1-sample Anderson–Darling goodness-of-fit test statistic for the hypothesis that $F(Y)$ in cases is uniformly distributed. This approach is in fact a special case within our proposed framework of comparing standardized marker distributions. In our opinion, applying a modified 2-sample version of the corresponding test directly to the standardized marker values is conceptually more straightforward.

The concept of covariate adjustment has only recently been developed for ROC analysis. The use of the covariate-specific percentiles provides a simple intuitive and easily implemented approach to adjust for covariates. Interestingly, arguments similar to (5.1) prove that the distribution of the covariate-specific case placement values, $1 - Q/100$, is the covariate-adjusted ROC curve, $\mathcal{A}\text{ROC}(t)$, proposed by Janes and Pepe (2008a,b,c). Thus, our methods for comparing distributions of the covariate-specific percentiles can be interpreted as methods to compare the $\mathcal{A}$ROC curves. Formal methods for comparing the $\mathcal{A}$ROC curves

have not been available heretofore. Our methods based on mean covariate-specific percentiles compare areas under the $\mathcal{A}$ROC curves while methods based on ranks provide an alternative approach.

In this paper, we focus primarily on comparing the ROC curve across the entire range of FPRs $\in (0, 1)$. In practice, one might focus on a part of the ROC curve that is of primary interest. For example, in screening studies, FPRs must be kept very low and so the ROC curve over a restricted range of FPR may be of interest. The percentile value framework is well suited to evaluations over restricted regions. If FPR is fixed at $u$, as we have shown, comparing $\mathrm{ROC}(u)$ can be achieved by comparing $\sum_{i=1}^{n_D} I(1 - \hat{Q}_i/100 \leqslant u)/n_D$ between samples. If FPR in the range $(0, u)$ is of interest, the partial AUC defined as $p\mathrm{AUC}(u) = \int_0^u \mathrm{ROC}(t)\mathrm{d}t$ has been proposed as the basis for marker comparisons (McClish, 1989). The empirical estimator written in terms of percentile values is

$$\widehat{p\mathrm{AUC}} = \frac{1}{n_D}\sum_{i=1}^{n_D} \max(\hat{Q}_i/100 - (1-u), 0),$$

where the empirical $\hat{F}$ is used to calculate $\hat{Q}_i$. This result follows by noting that

$$\begin{aligned} E[\max\{Q/100 - (1-u), 0\}] &= u - E\{\min(1 - Q/100, u)\} \\ &= u - \int_0^u t\,\mathrm{dROC}(t) - u\{1 - \mathrm{ROC}(u)\} \\ &= u - t\mathrm{ROC}(t)|_0^u + \int_0^u \mathrm{ROC}(t)\mathrm{d}t - u + u\mathrm{ROC}(u) \\ &= p\mathrm{AUC}(u). \end{aligned}$$

Returning now to the ovarian cancer and pancreatic cancer examples, suppose we are interested in comparisons based on $\mathrm{ROC}(u)$ and $p\mathrm{AUC}(u)$ for $u = 0.20$. We model the reference distributions parametrically and rely on the resampling variance for inference. In the ovarian cancer example, before covariate adjustment, $\widehat{\mathrm{ROC}}_1(u) = 0.5$, $\widehat{\mathrm{ROC}}_2(u) = 0.86$, with a difference of $-0.36$ (95% CI $= (-0.61, -0.12)$); $\widehat{p\mathrm{AUC}}_1(u) = 0.07$, $\widehat{p\mathrm{AUC}}_2(u) = 0.17$, with a difference of $-0.09$ (95% CI $= (-0.14, -0.05)$). After covariate adjustment, $\widehat{\mathrm{ROC}}_1(u) = 0.29$, $\widehat{\mathrm{ROC}}_2(u) = 0.83$, with a difference of $-0.54$ (95% CI $= (-0.80, -0.29)$); $\widehat{p\mathrm{AUC}}_1(u) = 0.04$, $\widehat{p\mathrm{AUC}}_2(u) = 0.16$, with a difference of $-0.12$ (95% CI $= (-0.16, -0.07)$). In the pancreatic cancer example, before covariate adjustment, $\widehat{\mathrm{ROC}}_1(u) = 0.79$, $\widehat{\mathrm{ROC}}_2(u) = 0.49$, with a difference of 0.3 (95% CI $= (0.11, 0.49)$); $\widehat{p\mathrm{AUC}}_1(u) = 0.14$, $\widehat{p\mathrm{AUC}}_2(u) = 0.06$, with a difference of 0.08 (95% CI $= (0.05, 0.12)$). After covariate adjustment, $\widehat{\mathrm{ROC}}_1(u) = 0.83$, $\widehat{\mathrm{ROC}}_2(u) = 0.3$, with a difference of 0.53 (95% CI $= (0.36, 0.71)$); $\widehat{p\mathrm{AUC}}_1(u) = 0.15$, $\widehat{p\mathrm{AUC}}_2(u) = 0.03$, with a difference of 0.12 (95% CI $= (0.08, 0.15)$). Comparisons based on points and partial areas under the curve agree with those based on the whole curve.

## 6. CONCLUDING REMARKS

Standardizing a biomarker or diagnostic test to a reference population of controls is not an entirely new concept. However, it is not yet a standard approach to biomarker evaluation. We suspect 2 reasons. First, ROC analysis has become the standard of practice (Baker, 2003), and second, formal methods have not been available for statistical inference that properly take account of sampling variability in the reference distribution. This paper provides remedies by providing methods for statistical inference and by noting that the approach is interchangeable with ROC analysis. We feel that the approach should be encouraged because of its conceptual simplicity.

The approach also opens up new avenues for evaluating biomarkers and diagnostic tests. For example, covariate adjustment is naturally handled within this framework. We illustrated that covariate adjustment can be important when comparing biomarkers or for comparing the performance of a biomarker in 2 populations. Pepe and Cai (2004) and Cai (2004) already showed how ROC regression can be accomplished by performing regression analysis of case standardized marker values. In the context of evaluating biomarkers for event time outcomes, one might use the risk set at time $t$ to standardize the biomarker for the subject that fails at $t$ (the case). Interestingly, it can be shown that the distribution of such standardized values is closely related to the time-dependent ROC curves developed by Heagerty and Zheng (2005). We hope that the methods presented here will encourage use of the percentile value standardized approach in practice and encourage further development of new techniques for biomarker evaluation.

## Acknowledgments

We thank Dr John A. Wellner for helpful comments and Dr Martin W. McIntosh for providing the ovarian cancer data. *Conflict of Interest:* None declared.

## Funding

National Institutes of Health (GM-54438 and CA-86368); Pacific Ovarian Cancer Research Consortium/SPORE in Ovarian Cancer (P50 CA83636, N.U.).

## References

Baker, S. G. (2003). The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute* **95**, 511–515.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 387–415.

Berres, M., Zehnder, A., Blasi, S. and Monsch, A. U. (2008). Evaluation of diagnostic scores with adjustment for covariates. *Statistics in Medicine* **27**, 1777–1790.

Cai, T. (2004). Semi-parametric ROC regression analysis with placement values. *Biostatistics* **5**, 45–60.

Cole, T. J. (1990). The LMS method for constructing normalized growth standards. *European Journal of Clinical Nutrition* **44**, 45–60.

Cuzick, J. (1985). A Wilcoxon-type test for trend. *Statistics in Medicine* **4**, 87–90.

Delong, E. R., Delong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics* **23**, 277–281.

Frischancho, A. R. (1990). *Anthropometric Standards for the Assessment of Growth and Nutritional Status*. Ann Arbor, MI: University of Michigan Press.

Hanley, J. A. and Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimate of the areas under receiver operating characteristic curves: an update. *Academic Radiology* **4**, 49–58.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

Janes, H. and Pepe, M. S. (2008a). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *UW Biostatistics Working Paper Series.* Working Paper 283.

JANES, H. AND PEPE, M. S. (2008b). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *American Journal of Epidemiology* **168**, 89–97.

JANES, H. AND PEPE, M. S. (2008c). Matching in studies of classification accuracy: implications for bias, efficiency, and assessment of incremental value. *Biometrics* **64**, 1–9.

KRUSKAL, W. H. AND WALLIS, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**, 583–621.

LI, G., TIWARI, R. C. AND WELLS, M. T. (1996). Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association* **91**, 689–698.

MCCLISH, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190–195.

MCINTOSH, M. W., DRESCHER, C., KARLAN, B., SCHOLLER, N., URBAN, N., HELLSTROM, K. E. AND HELLSTROM, I. (2004). Combining CA 125 and SMR serum markers for diagnosis and early detection of ovarian carcinoma. *Gynecologic Oncology* **95**, 9–15.

NAKAS, C., YIANNOUTSOS, C. T., BOSCH, R. J. AND MOYSSIADIS, C. (2003). Assessment of diagnostic markers by goodness-of-fit tests. *Statistics in Medicine* **22**, 2503–2513.

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

PEPE, M. S. AND CAI, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics* **60**, 528–535.

PEPE, M. S. AND LONGTON, G. M. (2005). Standardizing markers to evaluate and compare their performances. *Epidemiology* **16**, 598–603.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

SILVERMAN, B. W. AND YOUNG, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika* **74**, 469–479.

SUKHATME, S. AND BEAM, C. A. (1994). Stratification in nonparametric ROC studies. *Biometrics* **50**, 149–163.

SWETS, J. A. (1986). Form of empirical ROC's in discrimination and diagnosis tasks: implications of theory and measurement of performance. *Psychological Bulletin* **99**, 181–198.

VENKATRAMAN, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics* **56**, 1134–1138.

VENKATRAMAN, E. S. AND BEGG, C. B. (1996) A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* **83**, 835–848.

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

[*Received September 13, 2007; first revision January 14, 2008; second revision July 7, 2008; accepted for publication July 30, 2008*]