4-27-2007

Biomarker Evaluation Using the Controls as a Reference Population

Ying Huang
University of Washington, ying@u.washington.edu

Margaret Pepe
University of Washington, Fred Hutchinson Cancer Research Center, mspepe@u.washington.edu

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 ROC curve is typically used to describe the discriminatory capacity of a marker. However, for most statisticians, their familiarity with ROC methodology is limited. 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.

Two specific problems are considered here. The first problem 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. If so, failure to distinguish benign tumors from ovarian cancer limits the utility of this marker for both diagnostic and screening purposes. Let Y be the CA-125 measurement. Previously published data shown in Figure 1 are comprised of $\{Y_{\bar{D}i}, i = 1, ..., n_{\bar{D}}\}$ for controls, $\{Y_{1j}, j = 1, ..., n_1\}$ for cases with benign tumors, and $\{Y_{2j}, j = 1, ..., n_2\}$ 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 et al.,2004).

The second problem is to compare the discriminatory performances of two 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 et al., 1989), the biomarkers denoted by (Y_1, Y_2) are measured. The data are represented as $\{(Y_{1\bar{D}i}, Y_{2\bar{D}i}), i = 1, ..., n_{\bar{D}},\ (Y_{1Dj}, Y_{2Dj}), j = 1, ..., n_D\}$.

In this report, we start by setting these two 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.

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

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

This sort of standardization using a reference distribution is already commonplace in laboratory medicine and in 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.

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 of the normal reference range. A good biomarker would flag most cases as being outside of 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 tumors and malignant cancers in regards to their 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. An advantage of the standardization is that it 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 on the marker between disease status and disease type, we need only do a simple two sample comparison of Q between benign tumor cases and malignant cancer cases.

To compare two 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 two markers on a common scale where they can be compared using simple paired comparisons.

Adopting the control distribution as a reference to standardize a biomarker seems like a natural useful procedure. The approach has been taken in some biomarker studies (Frischancho, 1990; McIntosh et al., 2004), but it has never been formalized as a valid statistical method. Moreover, since in practice only a finite sample of controls are available, formal statistical procedures need to acknowledge sampling variability in the reference distribution. We will develop formal methods for inference here.

We can estimate F either empirically or parametrically with control data $\{Y_{\bar{D}i},\ i = 1, ..., n_{\bar{D}}\}$. Write $\hat{F}$ for the estimator which in the case of parametric estimation can also be written $F_{\hat{\theta}}$ where $\hat{\theta}$ is the estimated parameter for the model F_θ. 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*

3.1.1 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 $n_{\bar{D}}, n_1, n_2$ be the numbers of subjects in the control group and the 1^{st} and 2^{nd} case groups respectively. The next theorem is proved in supplementary appendices.

Theorem 1. Suppose marker observations are sampled independently and $n_{\bar{D}} \to \infty$, $\frac{n_1}{n_{\bar{D}}} \to \lambda_1 \in (0,1)$, $\frac{n_2}{n_{\bar{D}}} \to \lambda_2 \in (0,1)$, then $\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right)$ converges to a mean 0 normal random variable with variance σ^2, with

(a) $\sigma^2 = \text{var}\left(R_1(Y_{\bar{D}}) - R_2(Y_{\bar{D}})\right) + \frac{\text{var}(Q_1)}{\lambda_1} + \frac{\text{var}(Q_2)}{\lambda_2}$ if $\hat{F}$ is the empirical CDF, where $R_z(Y_{\bar{D}}) = P\left(Y_z < Y_{\bar{D}}\right)$ denotes the percentile value of the marker $Y_{\bar{D}}$ from a control within the z^{th} case distribution, and

(b) $\sigma^2 = \left(\frac{\partial\Delta}{\partial\theta}\right)^T \Sigma(\theta)\frac{\partial\Delta}{\partial\theta} + \frac{\text{var}(Q_1)}{\lambda_1} + \frac{\text{var}(Q_2)}{\lambda_2}$ if F is modeled parametrically, where $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}\left(\hat{\theta} - \theta\right)$ and we assume that Δ is differentiable with respect to θ. Thus the variability of $\hat{\Delta}$ comes from two sources, that due to sampling controls that form the reference distribution, and that due to sampling cases and calculating their percentile values given the reference distribution. In practice, we can use $\widehat{\text{var}}(\hat{R}_1(Y_{\bar{D}}) - \hat{R}_2(Y_{\bar{D}})) + \frac{\widehat{\text{var}}(\hat{Q}_1)}{n_1/n_{\bar{D}}} + \frac{\widehat{\text{var}}(\hat{Q}_2)}{n_2/n_{\bar{D}}}$ or $\left(\frac{\partial\hat{\Delta}}{\partial\theta}\right)|^T_{\theta=\hat{\theta}}\Sigma(\hat{\theta})\frac{\partial\hat{\Delta}}{\partial\theta}|_{\theta=\hat{\theta}} + \frac{\widehat{\text{var}}(Q_1)}{n_1/n_{\bar{D}}} + \frac{\widehat{\text{var}}(Q_2)}{n_2/n_{\bar{D}}}$ to consistently estimate σ^2 (where $\widehat{\text{var}}$ indicates the sample variance estimate). Another way to estimate σ^2 is to bootstrap, resampling subjects from the control and each case group separately. By calculating the variance of $\hat{\Delta} - \Delta$, we can construct a confidence interval for Δ, and formally test for equality of $E(Q_1)$ and $E(Q_2)$.

In the ovarian cancer study (McIntosh et al., 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

three 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 also computed the percentile values of CA-125 in each of the case groups, using either the empirical control distribution or under the assumption that log(CA-125) in controls follows a normal distribution after box-cox transformation. Distributions of the estimated percentile values are displayed by case group in Figure 1. Women with ovarian cancer appear to have larger percentile values of CA-125 compared to women with benign tumors, suggesting it better discriminates ovarian cancer than benign tumor from healthy women.

Let Q_1 and Q_2 be percentile values for women with benign tumors and women with ovarian cancer, respectively. We calculated 95% CI for Δ, the expected difference in mean percentile values between the two case groups. 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 Δ 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 Δ is $(-41.48, -9.46)$ based on the asymptotic variance, and $(-41.39, -9.56)$ based on the bootstrap variance. Inference based on the asymptotic and bootstrap variance agree fairly well here. The p-value for comparing $E(Q_1)$ and $E(Q_2)$ (denoted as "unconditional") is presented in Table 1. The population mean percentile values are significantly different (at $\alpha = 0.01$ level) between the two case groups, regardless of how we model the marker distribution in controls. The ability of CA-125 to identify ovarian cancer seems to be much better than is its ability to detect benign tumors.

3.1.2 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 two-sample tests for independent samples to be applied to compare case groups. The following formal proposition is proved in supplementary appendices.

Proposition 1. Under $H_0 : Q_1 =^d Q_2$, if the support of the marker Y in each case group is covered by its support in controls, then $Y_1 =^d Y_2$ 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 and Q_2 have the same distribution. That is, equal discriminatory performance is rejected. A common way to test the equality of distributions is to test for equality of means. Earlier we used the unconditional test to compare the means of Q_1 and Q_2. In another 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 $E(\hat{\Delta}|Y_i, i = 1, ..., n_{\bar{D}}) = 0$.

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

$$\begin{aligned} & \text{var}\left(\frac{1}{n_1}\sum_{j=1}^{n_1} \hat{Q}_{1j} - \frac{1}{n_2}\sum_{i=1}^{n_2} \hat{Q}_{2j} \middle| Y_i, i = 1, ..., n_{\bar{D}} \right) \\ = & \frac{\text{var}(\hat{Q}_1|Y_i, i = 1, ..., n_{\bar{D}})}{n_1} + \frac{\text{var}(\hat{Q}_2|Y_i, i = 1, ..., n_{\bar{D}})}{n_2} \end{aligned}$$

which can be consistently estimated by:

$$\frac{\widehat{\text{var}}(\hat{Q}_1)}{n_1} + \frac{\widehat{\text{var}}(\hat{Q}_2)}{n_2}.$$

On the other hand, the unconditional variance of $\hat{\Delta}$ can be estimated by:

$$\begin{aligned} & \frac{\widehat{\text{var}}(\hat{Q}_1)}{n_1} + \frac{\widehat{\text{var}}(\hat{Q}_2)}{n_2} + \frac{\widehat{\text{var}}(\hat{R}_1(Y_{\bar{D}}) - \hat{R}_2(Y_{\bar{D}}))}{n_{\bar{D}}} \\ \geq & \frac{\widehat{\text{var}}(\hat{Q}_1)}{n_1} + \frac{\widehat{\text{var}}(\hat{Q}_2)}{n_2} \end{aligned}$$

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 results in the top row of Table 1.

According to Proposition 1, $Q_1 =^d Q_2 \Leftrightarrow Y_1 =^d Y_2$. Therefore, an alternative way to test $H_0 : Q_1 =^d Q_2$ is to compare the distributions of Y_1 and Y_2, that is, the distributions of raw marker measurements between the two case groups. Standard two-sample tests for comparing two 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 tests based on percentile values have the same type-I error under the null hypothesis $H_0 : Y_1 =^d Y_2$ or $H_0 : Q_1 =^d Q_2$, but their powers may differ under different alternative hypotheses. In Table 1, we note that the test for equal means of Y_1 and Y_2 reaches the same conclusion as that for equal means of $\hat{Q}_1$ and $\hat{Q}_2$. This might not be true in other circumstances, depending on the particular control sample used as the reference. We do not include detailed illustrations here.

In summary, to compare a marker's ability to differentiate two different case groups from the same control group, we can compare means of their percentile values Q_1 and Q_2. On the one hand, if we are interested in constructing a confidence interval for $E(Q_1) - E(Q_2)$, we need to use unconditional inference that incorporates variability in controls as well as cases. We derived corresponding variance expressions. On the other hand, if our objective is simply to perform a hypothesis test for equality of the distributions of Q_1 and Q_2, we should use conditional methods because of their enhanced power.

3.2 *Rank Statistics*

The previous section deals 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 1(b) and 1(c)), the comparison between means does not tell the whole story about the difference between distributions. This motivates comparing distributions of percentile

values using other types of test statistics that are not based on means. For example, we can use rank-based statistics. The Wilcoxon rank sum test is often used for comparing two groups of independent samples. For the problem at hand, we need to acknowledge the correlation among $\hat{Q}'s$ when applying the Wilcoxon rank sum test to them.

In analogy with methods in the previous section, we can compare the ranks of $\hat{Q}_1$ and $\hat{Q}_2$ "unconditionally" or "conditionally". Applying the Wilcoxon rank sum test unconditionally to $\hat{Q}_1$ and $\hat{Q}_2$, the null hypothesis tested is $P(\hat{Q}_1 > \hat{Q}_2) = P(\hat{Q}_1 < \hat{Q}_2)$, which holds if $Q_1 =^d Q_2$ according to Proposition 1. Comparing the ranks of $\hat{Q}_1$ and $\hat{Q}_2$ conditional on the control sample, the null hypothesis tested is $P(\hat{Q}_1 > \hat{Q}_2 | Y_i, i = 1, ..., n_{\bar{D}}) = P(\hat{Q}_1 < \hat{Q}_2 | Y_i, i = 1, ..., n_{\bar{D}})$, which holds for all sets of control samples if $Q_1 =^d Q_2$. With conditional testing, the observations are independent and so standard Wilcoxon rank sum test can be applied. However, for the unconditional test, the variance of the Wilcoxon rank sum test statistic must be estimated using the bootstrap, resampling from controls and each case group.

In Table 1, both the conditional and unconditional Wilcoxon rank sum tests suggest significant differences in the distributions of CA-125 percentile values between benign tumor cases and ovarian cancer cases. Again, the conditional Wilcoxon rank sum test applied to $\hat{Q}$ 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 the raw marker measurements between the two case groups to test the null hypothesis $Q_1 =^d Q_2$. Consider the comparison between the Wilcoxon rank sum test applied to Y_1 and Y_2 and that applied to $\hat{Q}_1$ and $\hat{Q}_2$. If the transformation from Y to the corresponding $\hat{Q}$ does not change each observation's rank in the sample, then the rank based statistic is invariant to this transformation. This happens when F the marker distribution in controls is modeled

as a parametric family with strictly monotone increasing distribution function, but does not necessarily happen when F is estimated nonparametrically. In the latter case, the Wilcoxon rank sum test applied to $\hat{Q}_1, \hat{Q}_2$ can be different from that applied to Y_1, Y_2. Depending on their placement with respect to the control sample, the two groups of case marker measurement Y may be "squeezed" differently by the transformation to $\hat{Q}$. In particular, more ties may be created with the empirical CDF transformation. The increase in the number of ties will (1) potentially affect the value of the Wilcoxon rank sum test statistic (depending on how many pairs of $Y_{1i} > Y_{2j}$ or $Y_{1i} < Y_{2j}$ lead to $\hat{Q}_{1j} = \hat{Q}_{2j}$) and (2) reduce the variance of the test statistic. In the ovarian cancer data, the Wilcoxon rank sum test applied to $\hat{Q}'s$ and that applied to $Y's$ lead to the same conclusion (Table 1), but there are situations when different conclusions can be reached (results not shown).

We note that the variance of the Wilcoxon rank sum test statistic gets smaller as the number of ties in the data increases. Using the nonparametric bootstrap tends to create ties in the marker sample. Then when calculating percentile values, ties are created as a result of ties in the control and case samples when F is estimated empirically, or as a result of ties in case samples when F is estimated parametrically. This increase in ties due to sampling with replacement has the potential to lead to under-estimation of the variance. The severity of this problem depends on the sample size and the distribution of percentile values. We found in limited simulation studies that for small sample sizes and good classification accuracy, applying the Wilcoxon rank sum test with nonparametric bootstrap to $\hat{Q}$ can lead to anti-conservative type-I error, especially when F is estimated nonparametrically. A solution to this problem 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. We explored several bandwidths in simulation studies and found that the bandwidth that covers

around 40% of the total sample works reasonably well.

In summary, to compare the discriminatory performance of a marker across different case groups using rank based tests, we recommend (1) transforming the data to the estimated percentile values because the resulting test based on $\hat{Q}$ is more relevant to differences in diagnostic accuracies than differences on the raw marker distribution scale, and (2) using the conditional rather than unconditional Wilcoxon rank sum test because the conditional test can be performed with standard statistical software and is more powerful, whereas the unconditional test calls for smoothed bootstrap for variance estimation and does not have a sound theoretic basis for bandwidth selection.

3.3 *Adjusting for Covariates*

Suppose the biomarker distribution in controls varies with a covariate X, then the appropriate reference distribution should depend on X. We define the covariate specific percentile value

$$Q_X = 100 \times F(Y|X)$$

where $F(Y|X)$ is the cumulative distribution function of the marker in the control population with covariate value X. This is standard practice in clinical medicine for anthropometric measurements. For example, the percentiles of height for children are age and gender specific because these factors affect height in normal healthy children.

To compare women with benign tumor and women with ovarian cancer, we can evaluate the covariate specific percentiles values for each case group and compare them using two-sample statistics based on sample means or ranks as developed in section 3.1 and 3.2. 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 tumor, 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 a lot of controls per X category, a nonparametric approach to estimating $F(Y|X)$ can be taken. Otherwise a parametric model is employed. With $z = 1, 2$, 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 $n_{\bar{D}k}$ be the number of controls, and n_{zk} be the number of z^{th} type of cases in the k^{th} covariate category, $k = 1, ..., K$.

Theorem 2. Suppose $n_{\bar{D}} \to \infty, \frac{n_1}{n_{\bar{D}}} \to \lambda_1 \in (0,1), \frac{n_2}{n_{\bar{D}}} \to \lambda_2 \in (0,1)$. Suppose when X is discrete, $\frac{n_{\bar{D}k}}{n_{\bar{D}}} \to p_{\bar{D}k} \in (0,1)$, $\frac{n_{1k}}{n_1} \to p_{1k} \in (0,1)$, and $\frac{n_{2k}}{n_2} \to p_{2k} \in (0,1), k = 1, ..., K$. Then $\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right)$ converges to a mean 0 normal random variable with variance σ^2, where

(a) $\sigma^2 = \sum_k \left\{ \frac{\text{var}\left(R_1^k\left(Y_{\bar{D}}^k\right)\right)}{p_{\bar{D}k}/p_{1k}^2} + \frac{\text{var}\left(R_2^k\left(Y_{\bar{D}}^k\right)\right)}{p_{\bar{D}k}/p_{2k}^2} \right\} + \frac{\text{var}(Q_{1X})}{\lambda_1} + \frac{\text{var}(Q_{2X})}{\lambda_2}$ if $F(Y|X)$ is modeled with the empirical CDF, where $R_z^k\left(Y_{\bar{D}}^k\right) = P\left(Y_z^k < Y_{\bar{D}}^k\right)$ and the k superscript indicates cases and controls in covariate category k.

(b) $\sigma^2 = \frac{\partial \Delta}{\partial \theta}^T \Sigma(\theta) \frac{\partial \Delta}{\partial \theta} + \frac{\text{var}(Q_{1X})}{\lambda_1} + \frac{\text{var}(Q_{2X})}{\lambda_2}$ if $F(Y|X)$ is modeled parametrically, where $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}\left(\hat{\theta} - \theta\right)$ and we assume that Δ is differentiable with respect to θ and that $\mathcal{F} = \{F_\theta(y|x) : \theta \in \Theta\}$ is a Donsker class (Donsker, 1952). In practice, we can use $\sum_k \left\{ \frac{\widehat{\text{var}}\left(\hat{R}_1^k(Y_{\bar{D}}^k)\right)}{\frac{n_{\bar{D}k}}{n_{\bar{D}}}/\left(\frac{n_{1k}}{n_1}\right)^2} + \frac{\widehat{\text{var}}\left(\hat{R}_2^k(Y_{\bar{D}}^k)\right)}{\frac{n_{\bar{D}k}}{n_{\bar{D}}}/\left(\frac{n_{2k}}{n_2}\right)^2} \right\} + \frac{\widehat{\text{var}}(\hat{Q}_{1X})}{n_1/n_{\bar{D}}} + \frac{\widehat{\text{var}}(\hat{Q}_{2X})}{n_2/n_{\bar{D}}}$ or $\left(\frac{\partial \hat{\Delta}}{\partial \theta}\right)|_{\theta=\hat{\theta}}^T \Sigma(\hat{\theta}) \frac{\partial \hat{\Delta}}{\partial \theta}|_{\theta=\hat{\theta}} + \frac{\widehat{\text{var}}(\hat{Q}_{1X})}{n_1/n_{\bar{D}}} + \frac{\widehat{\text{var}}(\hat{Q}_{2X})}{n_2/n_{\bar{D}}}$ to consistently estimate σ^2.

To illustrate the comparison of covariate specific percentile values between two case groups, 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 $\mu = 10\times$ $log\left\{5 \times I(\text{benign tumors}) \times I(log(\text{CA-125}) > 2.2) + .8 \times I(\text{ovarian cancer}) + 1.5 \times log(\text{CA-125})\right\}$ and $\sigma = 4$. Figure 2 shows the distribution of log(CA-125) in healthy women, women with benign ovarian tumors, and women with ovarian cancer ignoring covariate X (marginal distributions, where F is modeled parametrically) or when X is equal to its $(.25, .5, .75)$ quantiles

in the whole sample. It appears that the distribution of log(CA-125) in controls varies with X. Moreover, the separations between controls and case groups differ with X. To calculate covariate-specific percentile values, we assume the distribution of log(CA-125) in controls conditional on a specific covariate value is normally distributed. The mean is modeled as a cubic B-spline in X, with pre-chosen knots at the $(.25, .5, .75)$ quantiles in the control sample.

Figure 3 plots the distributions of unadjusted and covariate specific percentile values of CA-125 for women with benign tumors and women with ovarian cancer. 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 adjusted percentile values have an approximately uniform distribution for women with benign tumors indicating that their distribution is the same as that for normal healthy 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 two groups of cases in regards to their covariate specific percentile values. All of the unconditional tests described in section 3.1 and 3.2 can be applied. P-values comparing $E(Q_{1X})$ and $E(Q_{2X})$ are 0.0002 based on the asymptotic variance and 0.0004 based on the bootstrap variance, while for the Wilcoxon rank sum test applied to $\hat{Q}_{1X}$ and $\hat{Q}_{2X}$ the p-value is less than 0.0001. All tests suggest that CA-125 has better discriminatory performance for identifying ovarian cancer compared to benign tumors. We also estimate the mean covariate specific percentile values for the two case groups and their difference Δ. As expected for benign tumors, $\overline{\hat{Q}}_{1X}$ is close to the uninformative marker value of 50 ($\overline{\hat{Q}}_{1X} = 50.13$), because their distribution is close to uniform (0,100). In the ovarian cancer group, $\overline{\hat{Q}}_{2X} = 88.10$ which is similar to the mean unadjusted percentile val-

ues $\overline{\hat{Q}}_2 = 90.17$. The difference in covariate adjusted means is $\hat{\Delta} = -37.96$, with 95% CI $(-57.76, -18.16)$ based on the asymptotic variance expression and $(-58.79, -17.13)$ based on the bootstrap variance. Observe that $\left|\overline{\hat{Q}}_{1X} - \overline{\hat{Q}}_{2X}\right|$ is larger than $\left|\overline{\hat{Q}}_1 - \overline{\hat{Q}}_2\right|$.

In summary, when the marker distribution in controls varies with a covariate, covariate specific percentile values can be calculated to eliminate potential confounding. Two 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.

4. Comparing Markers

Next consider the comparison of two markers with respect to their diagnostic accuracy. Suppose we have two types of subjects, cases and controls, with two markers measured on each subject. Let $n_D, n_{\bar{D}}$ be the number of cases and controls respectively. 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 (estimated) case percentile value for the z^{th} marker. 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 their percentile value standardized markers Q_1 and Q_2. If Q_1 tends to be larger than Q_2 then Q_1 is the better marker. Formally, let $\Delta = E(Q_1 - Q_2) = E(Q_1) - E(Q_2)$. The difference in sample means can serve as the basis of a test statistic $\hat{\Delta} = \overline{\hat{Q}}_1 - \overline{\hat{Q}}_2$.

In this two-marker setting, the correlation between estimated percentile values comes from two sources. First, two marker measurements measured on the same subject are correlated. Second for any particular marker, the estimated percentile values are correlated due to sampling variability in the controls used to calculate the reference distribution. We need to acknowledge this variation in making inference.

Theorem 3. Suppose $\frac{n_D}{n_{\bar{D}}} \to \lambda$ as $n_{\bar{D}} \to \infty$, then $\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right)$ converges to a mean 0 normal random variable with variance σ^2, where

(a) $\sigma^2 = \text{var}(R_1(Y_{1\bar{D}}) - R_2(Y_{2\bar{D}})) + \frac{\text{var}(Q_1 - Q_2)}{\lambda}$ if F_z is estimated with the empirical CDF, where $Y_{z\bar{D}}$ and Y_{zD} are the measurements of the z^{th} marker from 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 its case distribution (Delong et al., 1988),

(b) $\sigma^2 = \left(\frac{\partial \Delta}{\partial \theta}\right)^T \Sigma(\theta) \frac{\partial \Delta}{\partial \theta} + \frac{\text{var}(Q_1 - Q_2)}{\lambda}$ if F_z is modeled parametrically with parameter θ_z, $\theta = (\theta_1, \theta_2)$, $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}\left(\hat{\theta} - \theta\right)$ and we assume Δ is differentiable with respect to θ. In practice, σ^2 can be consistently estimated by $\widehat{\text{var}}(\hat{R}_1(Y_{1\bar{D}}) - \hat{R}_2(Y_{2\bar{D}})) + \frac{\widehat{\text{var}}(\hat{Q}_1 - \hat{Q}_2)}{n_D/n_{\bar{D}}}$ in (a) and $\left(\frac{\partial \hat{\Delta}}{\partial \theta}\right)^T \Sigma(\hat{\theta}) \frac{\partial \hat{\Delta}}{\partial \theta}|_{\theta=\hat{\theta}} + \frac{\widehat{\text{var}}(\hat{Q}_1 - \hat{Q}_2)}{n_D/n_{\bar{D}}}$ in (b). We could also bootstrap to estimate σ^2, resampling subjects from case and control groups separately.

Observe that, for this two-marker problem, the conditional test 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. That is, testing the null hypothesis that $\hat{Q}_1 | Y_{\bar{D}i}, i = 1, ..., n_{\bar{D}} =^d \hat{Q}_2 | Y_{\bar{D}i}, i = 1, ..., n_{\bar{D}}$ is not equivalent to testing the null hypothesis that $Q_1 =^d Q_2$.

The dataset we use for illustration here is from a pancreatic cancer serum biomarker study (Wieand et al., 1989). This is a case-control study including 90 cases with pancreatic cancer and 51 controls that had pancreatitis. 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 distribution of log(CA-19-9) and log(CA-125) for controls and cases separately. Also displayed are the distributions of the estimated percentile values in cases for each marker, with $F_z, z = 1, 2$ estimated empirically in Figure 5(b), and under the normal assumption after box-cox transformation in Figure 5(c). Clearly, the distribution of the estimated percentile values for CA-19-9 is shifted to the right compared

with CA-125, indicating that it is a better biomarker. In other words, more cases have high percentile values for CA-19-9 than for CA-125.

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 Δ is $(4.34, 26.73)$ using the asymptotic variance and similar, $(4.37, 26.70)$, 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 Δ is $(3.80, 26.15)$ using the asymptotic variance and $(3.57, 26.38)$ using the bootstrap variance. The differences are highly significant (Table 2). CA-19-9 is a better biomarker than CA-125 for pancreatic cancer.

In summary, to compare the diagnostic accuracies of two markers, we can use the controls to transform them to the percentile value scale and compare the percentile values in cases with the difference in 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. In particular, when $\hat{Q}_1$ and $\hat{Q}_2$ have similar expectations, a test comparing their means (section 4.1) will have low power. Rank based tests may be more powerful. Due to the complicated correlation structure, standard variance formulae for rank test statistics no longer apply. Instead, we use the bootstrap method to calculate their variances. Moreover, as discussed earlier, the conditional test is not applicable for the two marker problem. So only the unconditional test is discussed here. Heuristic proofs of the following three propositions are given in supplementary appendices.

Proposition 2.

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

Proposition 3.

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

Proposition 4

Let r_k be the rank of $\tilde{Q}_k$ where $\left\{\tilde{Q}_k, k = 1, ..., 2n_D\right\} = \left\{\hat{Q}_{1j}, j = 1, ..., n_D, \hat{Q}_{2j}, j = 1, ..., n_D\right\}$. Let $W = \sum_{k=1}^{n_D} r_k$ be the Wilcoxon rank sum test statistic. Then under $H_0 : Q_1 =^d Q_2$, $E(W) = \frac{n_D(2n_D+1)}{12}$ when F_z is estimated empirically.

We expect the corresponding results in propositions 2-4 to hold asymptotically when F_z is estimated parametrically.

Under $H_0 : Q_1 =^d Q_2$, or equivalently $\hat{Q}_1 =^d \hat{Q}_2$ (according to Proposition 2), the expectation of the Wilcoxon rank sum test statistic, W, applied to $\hat{Q}_1$ and $\hat{Q}_2$ is the same as $E(W)$ when W is applied to two groups of independent observations from the same distribution (Proposition 4); and the expectations of the Wilcoxon signed rank test statistic T and the Sign test statistic S applied to $\hat{Q}_1$ and $\hat{Q}_2$ is the same as $E(T)$ and $E(S)$ when those test statistics are applied to a paired sample where the two members in each pair have the same marginal distribution (Proposition 3). Therefore, to test $H_0 : Q_1 =^d Q_2$, we can use the rank based test statistics applied to $\hat{Q}_1$ and $\hat{Q}_2$, bootstrapping the variance. Here we face the same concern about under-estimation of the variance as in section 3.2. We use a smoothed bootstrap to minimize this problem. Asymptotic distribution theory appears to be very challenging. Table 2 displays p-values based on the rank tests for comparing the case distributions of CA-19-9 percentile values with CA-125 percentile values, using a bandwidth covering approximately 40% sample points in the smoothed bootstrap. All of these tests suggest a highly significant difference.

4.3 *Adjusting for Covariates*

We argued earlier that adjusting for covariates may be important when comparing two case groups. This is also potentially important when comparing two biomarkers. Suppose for example that biomarker values in the control group vary with study site in a multi-center 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 for reference (see the right side of Figure 5 for an example). Biomarker performance can appear to be worse than it is by using a pooled reference set. Markers may differ in regards to this phenomenon. Processing techniques that vary across sites may affect one marker but not another. 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 avoid such bias. In summary, covariate adjustment is required for covariates that affect the marker in controls. Note that pertinent covariates may be different for different markers.

Let $Q_{zX}(\hat{Q}_{zX})$ be the covariate specific percentile value (estimated) for the z^{th} marker, $z = 1, 2$, $\Delta = E(Q_{1X}) - E(Q_{2X})$, and $\hat{\Delta} = \overline{\hat{Q}_{1X}} - \overline{\hat{Q}_{2X}}$. When the covariate 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, $k = 1, ..., K$.

Theorem 4 Suppose $n_{\bar{D}} \to \infty, \frac{n_D}{n_{\bar{D}}} \to \lambda \in (0, 1)$, and for discrete covariate, $\frac{n_{\bar{D}k}}{n_{\bar{D}}} \to p_{\bar{D}k} \in (0, 1)$, $\frac{n_{Dk}}{n_D} \to p_{Dk} \in (0, 1), k = 1, ..., K$, then $\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right)$ converges to a mean 0 normal random variable with variance σ^2, where

(a) $\sigma^2 = \sum_k \frac{\text{var}(R_1^k(Y_{1\bar{D}}^k) - R_2^k(Y_{1\bar{D}}^k))}{p_{\bar{D}k}/p_{Dk}^2} + \frac{\text{var}(Q_{1X} - Q_{2X})}{\lambda}$ if $F(Y|X)$ is estimated empirically, 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 his covariate specific case distribution as the reference for z^{th} marker in the k^{th} covariate category,

(b) $\sigma^2 = \left(\frac{\partial \Delta}{\partial \theta}\right)^T \Sigma(\theta) \frac{\partial \Delta}{\partial \theta} + \frac{\text{var}(Q_{1X} - Q_{2X})}{\lambda}$ if $F(Y|X)$ is modeled parametrically for marker z with parameter estimate θ_z, $\theta = (\theta_1, \theta_2)$ and $\Sigma(\theta)$ is the asymptotic variance of $\sqrt{n_{\bar{D}}}\left(\hat{\theta} - \theta\right)$, we assume that Δ is differentiable with respect to θ and that $\mathcal{F} = \{F_\theta(y|x) : \theta \in \Theta\}$ is a Donsker class. In practice, σ^2 can be consistently estimated by $\sum_k \frac{\widehat{\text{var}}(\hat{R}_1^k(Y_{1\bar{D}}^k)) - \hat{R}_2^k(Y_{2\bar{D}}^k))}{\frac{n_{\bar{D}k}}{n_{\bar{D}}} / \left(\frac{n_{Dk}}{n_D}\right)^2} + \frac{\widehat{\text{var}}(\hat{Q}_{1X} - \hat{Q}_{2X})}{n_D / n_{\bar{D}}}$ in (a) and by $\left(\frac{\partial \hat{\Delta}}{\partial \theta}\right)|_{\theta=\hat{\theta}}^T \Sigma(\hat{\theta}_X) \frac{\partial \hat{\Delta}_X}{\partial \theta}|_{\theta=\hat{\theta}} + \frac{\widehat{\text{var}}(\hat{Q}_{1X} - \hat{Q}_{2X})}{n_D / n_{\bar{D}}}$ in (b).

To illustrate the use of adjusting for covariates when comparing markers, we simulate a discrete covariate X for the pancreatic cancer data. We set X to 1 for those with CA-125 above its median in the data, and 0 otherwise. 14 out of 51 (27.4%) controls and 57 out of 90 (63.3%) cases have covariate $X = 1$.

Figure 5 shows the probability distributions of log(CA-19-9) and log(CA-125) in control and case samples respectively within each covariate category. First we look at CA-19-9, it seems that the covariate does not have a dramatic influence on the reference control distribution. Thus covariate adjustment does not appear to be warranted for CA-19-9. On the other hand, covariate adjustment is warranted for CA-125. Within each covariate category, there is not much difference between cases and controls. However, since CA-125 is positively associated with the covariate and the case group has a higher percentage than controls of subjects with covariate $X = 1$, when data are pooled over covariate categories, the distribution of cases shifts to the right compared to the distribution of controls. In other words X is a confounder for the CA-125 marker.

Covariate specific percentile values for CA-19-9 ($\hat{Q}_{1X}$) and CA-125 ($\hat{Q}_{2X}$) in cases were calculated within each covariate category. Figure 6 plots the distributions. For CA-19-9, the case distribution of the estimated covariate specific percentile values is similar to that of the marginal percentile values, whereas for CA-125, covariate adjustment suggests poorer performance for CA-125 than the performance that ignores the covariate. That is, the confounding effect of X is removed by covariate adjustment.

When $F(Y|X)$ is estimated empirically for each marker, $\overline{\hat{Q}_{1X}} = 87.25$ for CA-19-9, and $\overline{\hat{Q}_{2X}} = 53.85$ for CA-125, with $\hat{\Delta} = 33.40$. The corresponding 95% CI for Δ is $(20.04, 46.76)$ using the asymptotic variance and $(20.83, 45.97)$ using the bootstrap variance. When $F(Y|X)$ is estimated parametrically for each marker, $\overline{\hat{Q}_{1X}} = 87.09$ for CA-19-9, and $\overline{\hat{Q}_{2X}} = 54.20$ for CA-125, with $\hat{\Delta} = 32.89$. The corresponding 95% CI for Δ is $(18.97, 46.81)$ using the asymptotic variance and $(20.38, 45.40)$ using the bootstrap variance. We compare $\hat{Q}_{1X}$ and $\hat{Q}_{2X}$ using mean and rank based tests as discussed in sections 4.1 and 4.2 (Table 2). 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 reference 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 (Frischancho, 1990; McIntosh et al., 2004). Interestingly it is equivalent to ROC analysis, which plays a central role in biomarker evaluation (Baker, 2003). The equivalence has been noted previously (Pepe and Cai, 2004; Pepe and Longton, 2005). In particular since the ROC curve, a plot of $TPR = P(Y > c|D = 1)$ versus $FPR = P(Y > c|D = 0)$, can be written as

$$\begin{aligned} ROC(t) &= P(Y > S^{-1}(t)|D = 1) \quad t \in (0, 1) \\ &= P(S(Y) < t|D = 1) \end{aligned}$$

where $S = 1 - F$, the ROC curve can be interpreted as the CDF of $1 - F(Y)$ in cases. Thus comparing case distributions of biomarker percentile values, $100 \times F(Y)$, is entirely

equivalent to comparing ROC curves. Empirical ROC curves for the ovarian and pancreatic cancer dataset are shown in Figure 7.

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, the average of case percentile values can be represented in terms of the area under the ROC curve (AUC)(Bamber, 1975),

$$AUC = E(Q)/100.$$

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 Theorem 2(a) are results for empirical AUC differences that have been previously reported (Sukhatme and Beam, 1994; Delong et al., 1988). However, their semi-parametric counterparts in Theorems 1(b) and 2(b) have not. Li et al. (1996) studied semi-parametric 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 distribution of control biomarker values. The Box-Cox family has precedent in modeling the reference distribution for anthropometric measures (Cole, 1990). Returning to the asymptotic results in Theorem 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 diagnostic accuracy of two tests using the area under the ROC curve is the lack of power to detect the difference in ROC curves when they have

the same area under the curve. As pointed out by Swets (1986), ROC curves are typically asymmetric, and two ROC curves with different asymmetries might cross each other but have the same AUC. Venkatraman and Begg (1996) developed a permutation test based procedure to compare two ROC curves with paired data. Extension of the permutation test to the case of continuous unpaired data was also proposed (Venkatraman, 2000). Extension to comparisons among more than two tests, however, might be computationally intensive.

The rank statistics described in sections 3.2 and 4.2 compare ROC curves as well. These can be interpreted as new ROC analysis techniques and provide an alternative way to compare ROC curves. On the other hand, 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).

The concept of covariate adjustment has only recently been developed for ROC analysis. The use of covariate specific percentiles provides a simple intuitive and easily implemented approach to adjust for covariates. Interestingly, arguments similar to those above prove that the distribution of covariate specific placement values, $1 - Q/100$, in cases, is the covariate adjusted ROC curve, $AROC(t)$, proposed by Janes and Pepe (2006, 2007). Thus, our methods for comparing distributions of covariate specific percentiles can be interpreted as methods to compare covariate adjusted ROC curves. Formal methods for comparing covariate adjusted ROC curves have not been available heretofore. Our methods based on mean covariate specific percentiles compare areas under the covariate adjusted ROC curves while methods based on ranks provide an alternative approach.

6. Concluding Remarks

Standardizing a biomarker or diagnostic test to a reference population of controls is not an entirely new concept (Frischancho, 1990; McIntosh et al. 2004). However it is not yet a

standard approach to biomarker evaluation. We suspect two 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 the approach is interchangeable with ROC analysis. We feel that the approach should be encouraged because of its conceptual simplicity, putting ROC analysis within mainstream familiar data analytic methods.

Equally important, the approach 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 in comparing the performance of a biomarker in two 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 of 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 recently 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.

7. Acknowledgments

The authors are grateful for support provided by NIH grants GM-54438 and CA-86368, and for funding from the Pacific Ovarian Cancer Research Consortium (POCRC)/SPORE in Ovarian Cancer (P50 CA83636, N.U.). And we thank Dr. Martin W. McIntosh for providing the ovarian cancer data.

References

Baker, S. G. (2003) The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *JNCI*, 95(7), 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.

Cai, T. (2004) Semi-parametric ROC regression analysis with placement values. *Biostatistics*, 5(1), 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., Clarke-Pearson, D.L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 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.

Li, G., Tiwari, R.C., and Wells, M.T. (1996) Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *JASA*, 91(434), 689-698.

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(1), 92-105.

Janes, H., and Pepe, M. S. (2006) Adjusting for covariate effects in biomarker studies using the subject specific threshold ROC curve. Submitted to *JASA*.

Janes, H., and Pepe, M. S. (2007) Matching in studies of classification accuracy: Implications

for bias, efficiency, and assessment of incremental value. Accepted by *Biometrics*.

Kruskal, W.H. and Wallis, W.A. (1952) Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* 47, 583-621.

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 to diagnosis and early detection of ovarian carcinoma. *Gynecologic Oncology*, 95, 9-15.

Pepe, M. S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series 28.

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(5), 598-603.

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

Silverman, B.W. and Young, G. A. (1987) The bootstrap: to smooth or not to smooth? *Biometrika* 74(3), 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. *Psychol. Bull* 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(4), 835-848.

Frischancho, A. R. (1990) Anthropometric standards for the assessment of growth and nutritional status. Ann Arbor: University of Michigan Press.

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, 585-92.

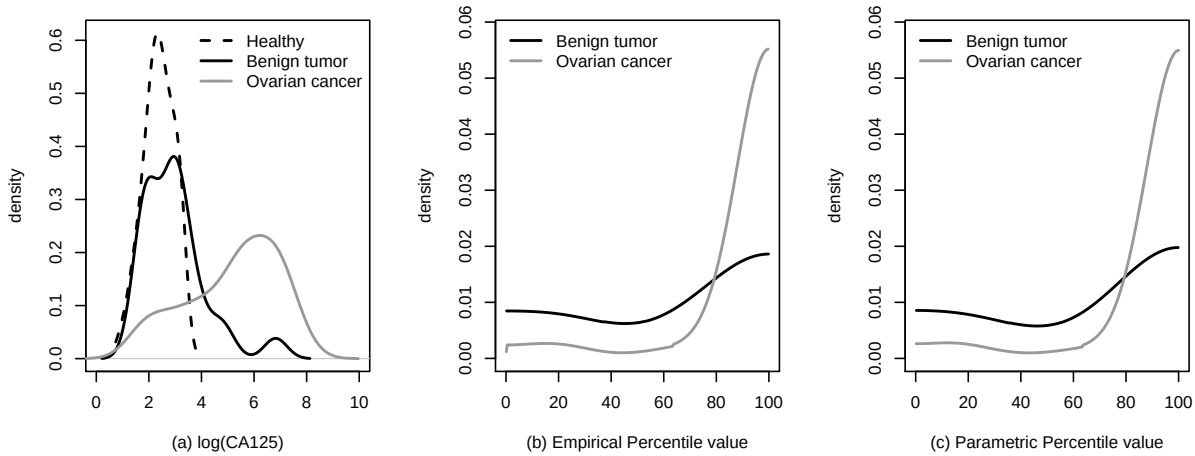

Figure 1. (a) Distributions of log(CA-125) in healthy women, women with benign ovarian tumors, and women with ovarian cancer. (b),(c) Distributions of estimated percentile values in benign tumor cases and ovarian cancer cases. Percentile values are calculated with the empirical distribution of CA-125 in controls in (b). The distribution of CA-125 in controls is assumed to be normal after a Box-Cox transformation in (c).

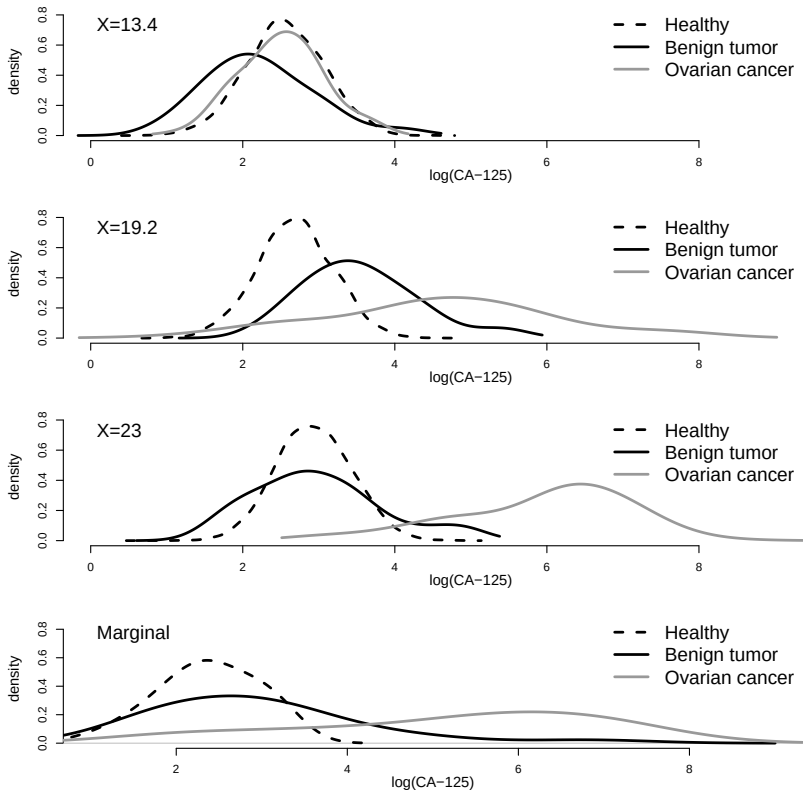

Figure 2. Distributions of log(CA-125) in healthy women, women with benign ovarian tumors, and women with ovarian cancer for specified covariate value and marginally.

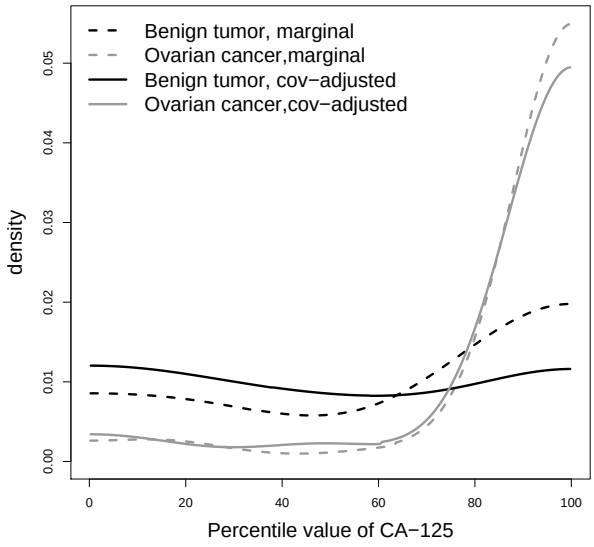

Figure 3. Distributions of estimated percentile values of CA-125 for women with benign ovarian tumors, and women with ovarian cancer. Here 'cov-adjusted' indicates that covariate specific control reference distributions were employed, while 'marginal' indicates that the entire set of controls were used as a single reference group..

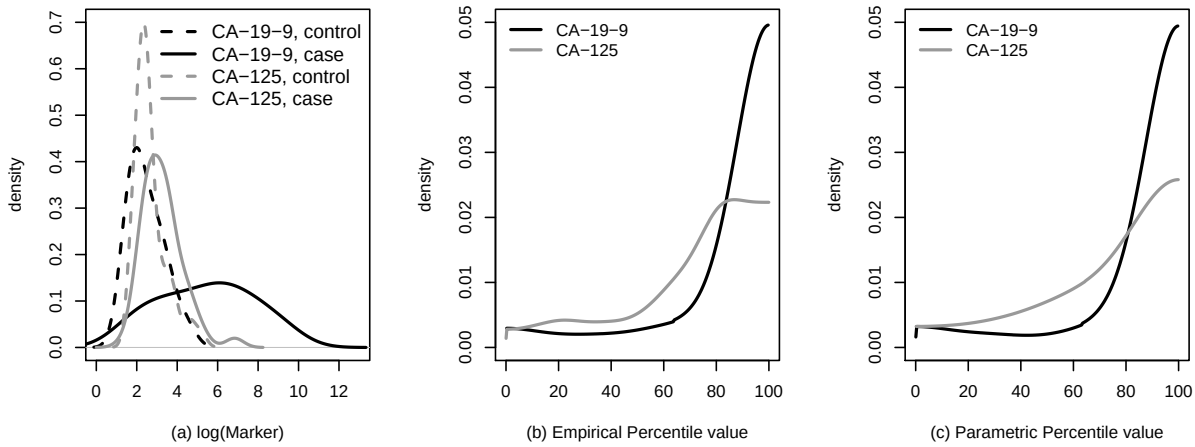

Figure 4. (a) Distributions of log(CA-19-9) and log(CA-125) in controls and cases, (b) distributions of estimated case percentile values when control distributions are estimated empirically, and (c) distributions of estimated case percentile values when control distributions are assumed to be normal after the Box-Cox transformation.

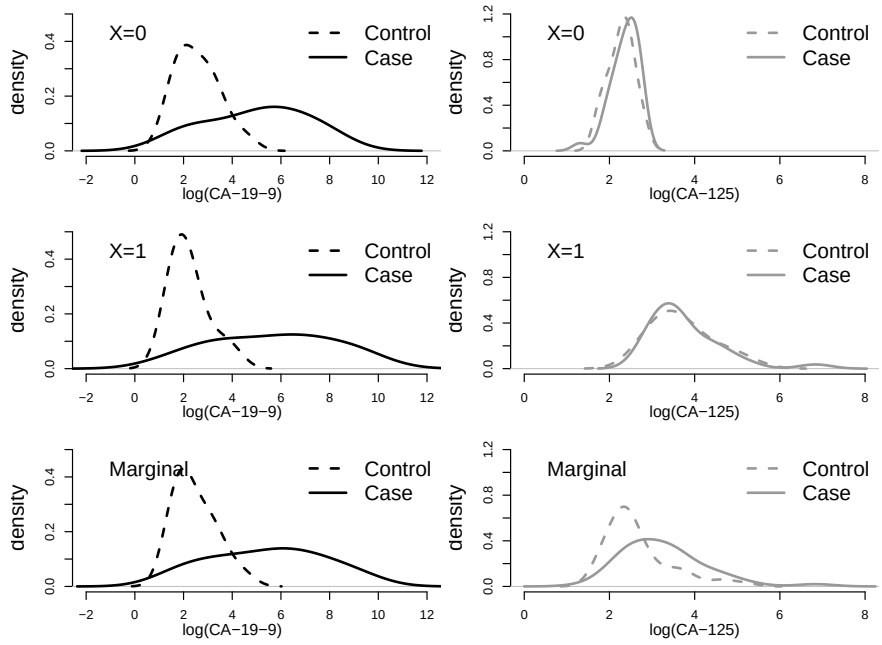

Figure 5. Distributions of log(CA-19-9) and log(CA-125) in controls and cases within each covariate category and marginally.

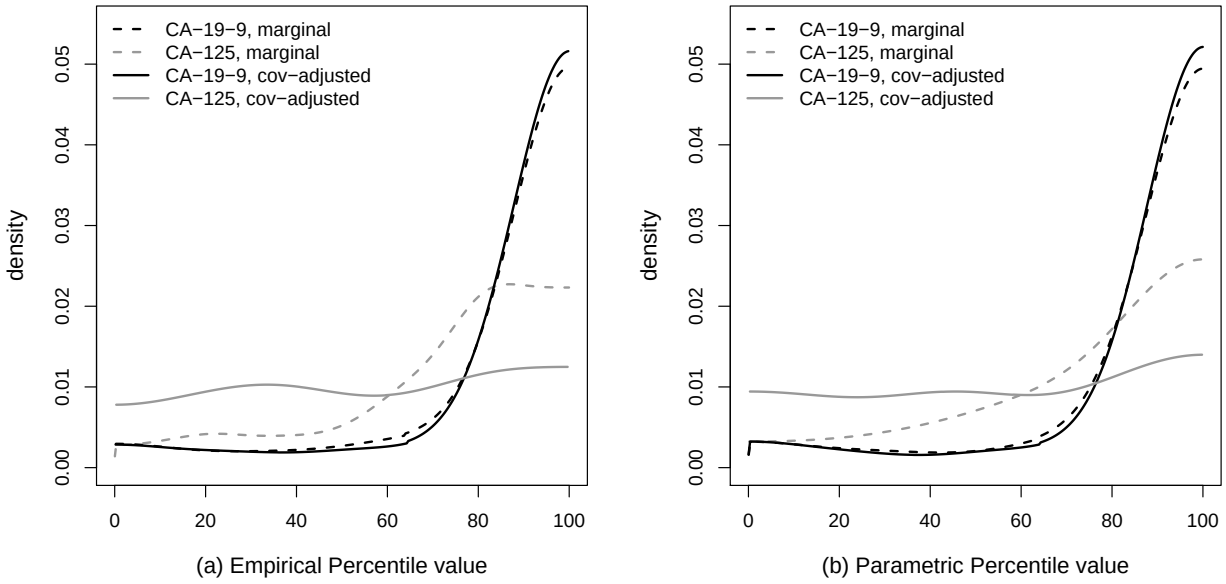

Figure 6. Distributions of estimated case percentile values of CA-19-9 and CA-125. Here 'cov-adjusted' indicates that covariate specific control reference distributions were employed, while 'marginal' indicates that the entire set of controls were used as a single reference group.

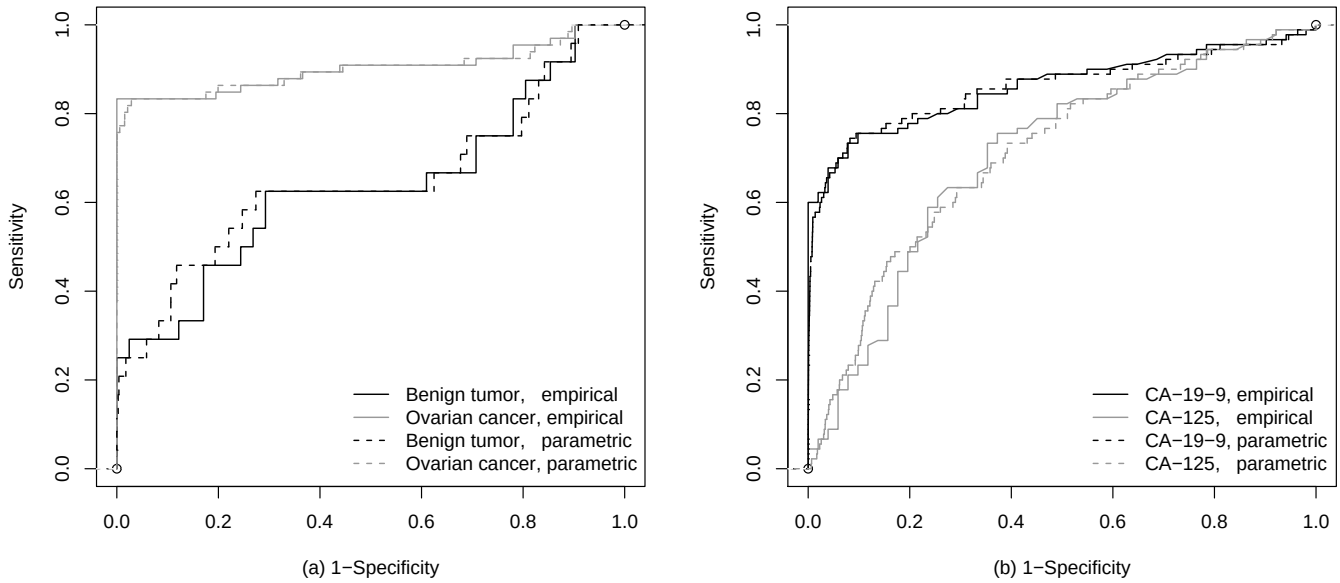

Figure 7. ROC curves (control distribution is empirically or parametrically estimated) for benign tumor cases and ovarian cancer cases in ovarian cancer data (a), and for CA-19-9 and CA-125 in pancreatic cancer data (b).

Table 1

p-value of tests comparing case percentile value distributions between benign tumor cases and ovarian cancer cases. $n_{\bar{D}} = 41, n_1 = 24, n_2 = 66$. *Tests comparing raw marker values between benign tumor cases and ovarian cancer cases yielded* $p < 0.0001$ *for both the t-test and the Wilcoxon rank sum test.*

	Unconditional				**Conditional**			
Test	$\hat{F}$ **empirical**		$\hat{F}$ **parametric**		$\hat{F}$ **empirical**		$\hat{F}$ **parametric**	
	Asym[1]	**Boot**[2]	**Asym**	**Boot**	**Asym**	**Boot**	**Asym**	**Boot**
Mean	0.0009	0.0006	0.0018	0.0013	0.0005	0.0003	0.0012	0.0009
Rank	-	$< 0.0001^s$	-	$< 0.0001^s$	< 0.0001	$< 0.0001^s$	< 0.0001	$< 0.0001^s$

Asym[1]: asymptotic variance
Boot[2]: nonparametric bootstrap variance or smoothed bootstrap variance (^{s})

Table 2

p-value for comparing percentile value distributions between CA-19-9 and CA-125. $n_{\bar{D}} = 51, n_D = 90$

Test Statistic	$\hat{F}$ **empirical CDF**		$\hat{F}$ **parametric**	
	Asym[1]	**Boot**[2]	**Asym**	**Boot**
		Marginal		
Mean Difference	0.007	0.007	0.009	0.01
WRS[3]	-	$< 0.0001^s$	-	0.0006^s
WSR[4]	-	$< 0.0001^s$	-	0.0001^s
Sign[5]	-	$< 0.0001^s$	-	$< 0.0001^s$
		Covariate Adjusted		
Mean Difference	< 0.0001	< 0.0001	< 0.0001	< 0.0001
WRS	-	< 0.0001	-	< 0.0001
WSR	-	< 0.0001	-	< 0.0001
Sign	-	< 0.0001	-	< 0.0001

Asym[1]: asymptotic variance
Boot[2]: nonparametric bootstrap variance, or smoothed bootstrap variance^s
WRS[3]: the Wilcoxon rank sum test statistic
WSR[4]: the Wilcoxn signed rank test statistic
Sign[5]: the Sign test statistic