Biostatistics (2007), **8**, 1, *pp.* 1–8
doi:10.1093/biostatistics/kxl005
Advance Access publication on May 15, 2006

Supplementary Material to "Biomarker Evaluation and Comparison Using the Controls as a Reference Population"

YING HUANG*

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

MARGARET SULLIVAN PEPE

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

WEB APPENDIX A: TABLES AND FIGURES

Table 1. *P-values for comparing the percentile value distributions of 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
t-test	0.0009	0.0006	0.0018	0.0013	0.0005	0.0003	0.0012	0.0009
WRS[3]	-	$< 0.0001^s$	-	$< 0.0001^s$	< 0.0001	$< 0.0001^s$	< 0.0001	$< 0.0001^s$

[1] asymptotic variance

[2] nonparametric bootstrap variance or smoothed bootstrap variance indicated by superscript ^{s}

[3] Wilcoxon rank sum test

Table 2. *P-values for comparing the case percentile value distributions of 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[1]	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

[1] asymptotic variance

[2] nonparametric bootstrap variance, or smoothed bootstrap variance indicated by superscript ^{s}

[3] Wilcoxon rank sum test statistic

[4] Wilcoxon signed rank test statistic

[5] Sign test statistic

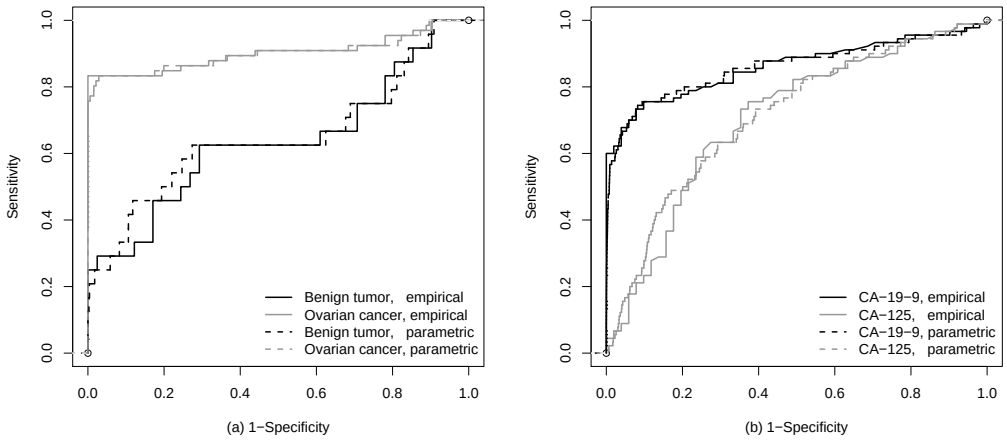

Fig. 1. A. ROC curves (control distribution is empirically or parametrically estimated) (a) for benign tumor cases and ovarian cancer cases in the ovarian cancer data, and (b) for CA-19-9 and CA-125 in the pancreatic cancer data.

WEB APPENDIX B: PROOF OF THEOREMS

0.1 *Proof of Theorem 1*

With $z = 1, 2$, let $Y_{\bar{D}i}, i = 1, ..., n_{\bar{D}}$ and $Y_{zj}, j = 1, ..., n_z$ be the marker measurements in controls and the z^{th} type of cases. Let F and G_z be the corresponding marker distribution functions. When F is estimated empirically,

$$\begin{aligned}
&\sqrt{n_{\bar{D}}}\left(\hat{\Delta}-\Delta\right)\\
=\ &\sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1 n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}}\sum_{j=1}^{n_1} I\left(Y_{\bar{D}i}\leqslant Y_{1j}\right)-\frac{1}{n_2 n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}}\sum_{j=1}^{n_2} I\left(Y_{\bar{D}i}\leqslant Y_{2j}\right)-P(Y_{\bar{D}}\leqslant Y_1)+P(Y_{\bar{D}}\leqslant Y_2)\right\}\times 100\\
=^{*}\ &\sqrt{n_{\bar{D}}}\left(\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}}\left[1-G_1(Y_{\bar{D}i})-\{1-G_2(Y_{\bar{D}i})\}\right]-P(Y_{\bar{D}}\leqslant Y_1)+P(Y_{\bar{D}}\leqslant Y_2)\right)\times 100\\
+\ &\sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1}\sum_{j=1}^{n_1}F(Y_{1j})-P(Y_{\bar{D}}\leqslant Y_1)\right\}\times 100-\sqrt{n_{\bar{D}}}\left\{\frac{1}{n_2}\sum_{j=1}^{n_2}F(Y_{2j})-P(Y_{\bar{D}}\leqslant Y_2)\right\}\times 100+o_p(1)\\
=\ &-\frac{1}{\sqrt{n_{\bar{D}}}}\sum_{i=1}^{n_{\bar{D}}}\{R_1(Y_{\bar{D}i})-R_2(Y_{\bar{D}i})-E(R_1-R_2)\}\\
+\ &\frac{1}{\sqrt{\lambda_1 n_1}}\sum_{j=1}^{n_1}\{Q_{1j}-E(Q_1)\}-\frac{1}{\sqrt{\lambda_2 n_2}}\sum_{j=1}^{n_2}\{Q_{2j}-E(Q_2)\}+o_p(1),
\end{aligned}$$

where * can be proved with the U-statistic theory (van der Vaart, 1998).

When F is modeled parametrically, assume that $\sqrt{n}\left(\hat{\theta}-\theta\right)$ can be represented as $\frac{1}{\sqrt{n_{\bar{D}}}}\sum_{i=1}^{n_{\bar{D}}}\psi_i + o_p(1)$, where $\psi_i, i = 1, ..., n_{\bar{D}}$ are independent identically distributed variables with $E(\psi_i) = 0$ and $\text{var}(\psi_i) = \Sigma(\theta)$, we have

$$\begin{aligned}
&\sqrt{n_{\bar{D}}}\left(\hat{\Delta}-\Delta\right)\\
=\ &\sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1}\sum_{j=1}^{n_1}F_{\hat{\theta}}(Y_{1j})-\frac{1}{n_2}\sum_{j=1}^{n_2}F_{\hat{\theta}}(Y_{2j})-E_{Y_1}F_{\theta}(Y)+E_{Y_2}F_{\theta}(Y)\right\}\times 100\\
=^{*}\ &\sqrt{n_{\bar{D}}}\left\{E_{Y_1}F_{\hat{\theta}}(Y)-E_{Y_1}F_{\theta}(Y)-E_{Y_2}F_{\hat{\theta}}(Y)+E_{Y_2}F_{\theta}(Y)\right\}\times 100\\
+\ &\sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1}\sum_{j=1}^{n_1}F_{\theta}(Y_{1j})-\frac{1}{n_2}\sum_{j=1}^{n_2}F_{\theta}(Y_{2j})-E_{Y_1}F_{\theta}(Y)+E_{Y_2}F_{\theta}(Y)\right\}\times 100+o_p(1)\\
=\ &\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}}\frac{\partial\Delta_i}{\partial\theta}\psi_i+\frac{1}{\sqrt{\lambda_1 n_1}}\sum_{j=1}^{n_1}\{Q_{1j}-E(Q_1)\}-\frac{1}{\sqrt{\lambda_2 n_2}}\sum_{j=1}^{n_2}\{Q_{2j}-E(Q_2)\}+o_p(1),
\end{aligned}$$

where * follows since $\mathcal{F}=\{F_\theta(y):\theta\in\Theta\}$ is a Donsker class (van der Vaart and Wellner, 1996, Theorem 2.7.5 on page 159).

0.2 *Proof of Theorem 2*

With $z = 1, 2$ and $k = 1, ..., K$, let $Y^k_{\bar{D}i}, i = 1, ..., n_{\bar{D}k}$ and $Y^k_{zj}, j = 1, ..., n_{zk}$ be the marker measurements for controls and the z^{th} type of cases in the k^{th} covariate category. Let F^k and G^k_z be the corresponding marker distribution functions. When F^k is estimated empirically,

$$
\begin{aligned}
&\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right) \\
&= \sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1}\sum_{k=1}^{K}\sum_{i=1}^{n_{\bar{D}k}}\sum_{j=1}^{n_{1k}}\frac{1}{n_{\bar{D}k}} I\left(Y^k_{\bar{D}i} \leqslant Y^k_{1j}\right) - \frac{1}{n_2}\sum_{k=1}^{K}\sum_{i=1}^{n_{\bar{D}}}\sum_{j=1}^{n_{2k}}\frac{1}{n_{\bar{D}k}} I\left(Y^k_{\bar{D}i} \leqslant Y^k_{2j}\right)\right. \\
&\quad \left. - \sum_{k=1}^{K} p_{1k} P\left(Y^k_{\bar{D}} \leqslant Y^k_1\right) + \sum_{k=1}^{K} p_{2k} P\left(Y^k_{\bar{D}} \leqslant Y^k_2\right)\right\} \times 100 \\
&= \sqrt{n_{\bar{D}}}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n_{\bar{D}k}}\frac{1}{n_{\bar{D}k}}\left[p_{1k}\left\{1 - G^k_1\left(Y^k_{\bar{D}i}\right)\right\} - p_{2k}\left\{1 - G^k_2\left(Y^k_{\bar{D}i}\right)\right\}\right] - \sum_{k=1}^{K} p_{1k} P\left(Y^k_{\bar{D}} \leqslant Y^k_1\right)\right. \\
&\quad \left. + \sum_{k=1}^{K} p_{2k} P\left(Y^k_{\bar{D}} \leqslant Y^k_2\right)\right\} \times 100 + \sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1}\sum_{k=1}^{K}\sum_{j=1}^{n_{1k}} F^k(Y^k_{1j}) - \sum_{k=1}^{K} p_{1k} P\left(Y^k_{\bar{D}} \leqslant Y^k_1\right)\right\} \times 100 \\
&\quad - \sqrt{n_{\bar{D}}}\left\{\frac{1}{n_2}\sum_{k=1}^{K}\sum_{j=1}^{n_2} F^k\left(Y^k_{2j}\right) - \sum_{k=1}^{K} p_{2k} P\left(Y^k_{\bar{D}} \leqslant Y^k_2\right)\right\} \times 100 + o_p(1) \\
&= \sum_{k=1}^{K} -\frac{1}{\sqrt{n_{\bar{D}k}}\, p_{\bar{D}k}} \sum_{i=1}^{n_{\bar{D}k}} \left\{p_{1k} R^k_1\left(Y^k_{\bar{D}}\right) - p_{2k} R^k_2\left(Y^k_{\bar{D}}\right) - p_{1k} E\left(R^k_1\right) + p_{2k} E\left(R^k_2\right)\right\} \\
&\quad + \frac{1}{\sqrt{\lambda_1 n_1}}\sum_{j=1}^{n_1}\left\{Q_{1Xj} - E\left(Q_{1X}\right)\right\} - \frac{1}{\sqrt{\lambda_2 n_2}}\sum_{j=1}^{n_2}\left\{Q_{2Xj} - E\left(Q_{2X}\right)\right\} + o_p(1).
\end{aligned}
$$

This proves Theorem 2(a). Now, consider when F is estimated parametrically and suppose the parameter θ can be represented as $\sqrt{n_{\bar{D}}}\left(\hat{\theta} - \theta\right) = \frac{1}{\sqrt{n_{\bar{D}}}}\sum_{i=1}^{n_{\bar{D}}}\psi_i + o_p(1)$,where $\psi_i, i = 1, ..., n_{\bar{D}}$ are independent identically distributed variables with $E(\psi_i) = 0$ and var$(\psi_i) = \Sigma(\theta)$. Let $X_{zj}, j = 1, ..., n_z$ be the covariate measurements for the z^{th} type of cases, we have

$$
\begin{aligned}
&\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right) \\
&= \sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1}\sum_{j=1}^{n_1} F_{\hat{\theta}}(Y_{1j}|X_{1j}) - \frac{1}{n_2}\sum_{j=1}^{n_2} F_{\hat{\theta}}(Y_{2j}|X_{2j}) - E_{Y_1,X_1} F_\theta(Y|X) + E_{Y_2,X_2} F_\theta(Y|X)\right\} \times 100 \\
&= \sqrt{n_{\bar{D}}}\left\{E_{Y_1,X_1} F_{\hat{\theta}}(Y|X) - E_{Y_1,X_1} F_\theta(Y|X) - E_{Y_2,X_2} F_{\hat{\theta}}(Y|X) + E_{Y_2,X_2} F_\theta(Y|X)\right\} \times 100 \\
&+ \sqrt{n_{\bar{D}}}\left\{\frac{1}{n_1}\sum_{j=1}^{n_1} F_\theta(Y_{1j}|X_{1j}) - \frac{1}{n_2}\sum_{j=1}^{n_2} F_\theta(Y_{2j}|X_{2j}) - E_{Y_1,X_1} F_\theta(Y|X) + E_{Y_2,X_2} F_\theta(Y|X)\right\} \times 100 + o_p(1) \\
&= \frac{1}{\sqrt{n_{\bar{D}}}}\sum_{i=1}^{n_{\bar{D}}}\frac{\partial \Delta_i}{\partial \theta}\psi_i + \frac{1}{\sqrt{\lambda_1 n_1}}\sum_{j=1}^{n_1}\{Q_{1Xj} - E(Q_{1X})\} - \frac{1}{\sqrt{\lambda_2 n_2}}\sum_{j=1}^{n_2}\{Q_{2Xj} - E(Q_{2X})\} + o_p(1),
\end{aligned}
$$

where $\{E_{Y_z,X_z}, z = 1, 2\}$ is integrated over the joint distribution of the marker Y and the covariate X in the z^{th} type of cases.

0.3 *Proof of Theorem 3*

The first result, 3(a) has been proved (Delong *and others*, 1988). For the second result, with $z = 1, 2$, let $Y_{zDj}, j = 1, ..., n_D$ be marker z measurements in cases. Let F_{θ_z} be the distribution function for marker z in controls and $\theta = (\theta_1, \theta_2)$. Suppose $\sqrt{n_{\bar{D}}}\left(\hat{\theta} - \theta\right)$ can be represented as $\frac{1}{\sqrt{n_{\bar{D}}}} \sum_{i=1}^{n_{\bar{D}}} \psi_i + o_p(1)$,where $\psi_i, i = 1, ..., n_{\bar{D}}$ are independent identically distributed variables with $E(\psi_i) = 0$ and var$(\psi_i) = \Sigma(\theta)$, we have

$$
\begin{aligned}
&\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right) \\
&= \sqrt{n_{\bar{D}}}\left[\frac{1}{n_D}\sum_{j=1}^{n_D}\left\{F_{\hat{\theta}}(Y_{1D_j}) - F_{\hat{\theta}}(Y_{2D_j})\right\} - E_{Y_{1D}}F_{\theta_1}(Y) + E_{Y_{2D}}F_{\theta_2}(Y)\right] \times 100 \\
&= \sqrt{n_{\bar{D}}}\left\{E_{Y_{1D}}F_{\hat{\theta}_1}(Y) - E_{Y_{1D}}F_{\theta_1}(Y) - E_{Y_{2D}}F_{\hat{\theta}_2}(Y) + E_{Y_{2D}}F_{\theta_2}(Y)\right\} \times 100 \\
&+ \sqrt{n_{\bar{D}}}\left[\frac{1}{n_D}\sum_{j=1}^{n_D}\left\{F_{\theta_1}(Y_{1D_j}) - F_{\theta_2}(Y_{2D_j})\right\} - E_{Y_{1D}}F_{\theta_1}(Y) + E_{Y_{2D}}F_{\theta_2}(Y)\right] \times 100 + o_p(1) \\
&= \frac{1}{\sqrt{n_{\bar{D}}}}\sum_{i=1}^{n_{\bar{D}}}\frac{\partial\Delta_i}{\partial\theta}\psi_i + \frac{1}{\sqrt{\lambda n_D}}\sum_{j=1}^{n_D}\{Q_{1j} - Q_{2j} - E(Q_1 - Q_2)\} + o_p(1)
\end{aligned}
$$

and result 3(b) follows.

0.4 *Proof of Theorem 4*

With $z = 1, 2$, let $Y^k_{z\bar{D}i}, i = 1, ..., n_{\bar{D}k}$ and $Y^k_{zDj}, j = 1, ..., n_{Dk}$ be the measurements of marker z in controls and cases respectively in the k^{th} covariate category. Let F^k_z and G^k_z be the corresponding distribution functions. When F^k is estimated empirically,

$$
\begin{aligned}
&\sqrt{n_{\bar{D}}}\left(\hat{\Delta} - \Delta\right) = \sqrt{n_{\bar{D}}}\left[\frac{1}{n_D}\sum_{k=1}^{K}\sum_{i=1}^{n_{\bar{D}k}}\sum_{j=1}^{n_{Dk}}\frac{1}{n_{\bar{D}k}}\left\{I\left(Y^k_{1\bar{D}i} \leqslant Y^k_{1D_j}\right) - I\left(Y^k_{2\bar{D}i} \leqslant Y^k_{2Dj}\right)\right\}\right. \\
&\left. - \sum_{k=1}^{K} p_{Dk}P\left(Y^k_{\bar{D}} \leqslant Y^k_{1D}\right) + \sum_{k=1}^{K} p_{Dk}P\left(Y^k_{2\bar{D}} \leqslant Y^k_{2D}\right)\right] \times 100
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{n_{\bar{D}}}\left(\sum_{k=1}^{K}\sum_{i=1}^{n_{\bar{D}k}}\frac{p_{Dk}}{n_{\bar{D}k}}\left[1-G_1^k\left(Y_{1\bar{D}i}^k\right)-\left\{1-G_2^k\left(Y_{2\bar{D}i}^k\right)\right\}\right]-\sum_{k=1}^{K}p_{Dk}P\left(Y_{1\bar{D}}^k\leqslant Y_{1D}^k\right)\right.\\
&\left.+\sum_{k=1}^{K}p_{Dk}P\left(Y_{2\bar{D}}^k\leqslant Y_{2D}^k\right)\right)\times 100+\sqrt{n_{\bar{D}}}\left[\frac{1}{n_D}\sum_{k=1}^{K}\sum_{j=1}^{n_{Dk}}\left\{F_1^k\left(Y_{1D_j}^k\right)-F_2^k(Y_{2D_j}^k\right)\right\}\right.\\
&\left.-\sum_{k=1}^{K}p_{Dk}\left\{P\left(Y_{1\bar{D}}^k\leqslant Y_{1D}^k\right)-P\left(Y_{2\bar{D}}^k\leqslant Y_{2D}^k\right)\right\}\right]\times 100+o_p(1)\\
&=\sum_{k=1}^{K}-\frac{1}{\sqrt{n_{\bar{D}k}}}\sum_{i=1}^{n_{\bar{D}k}}\frac{p_{Dk}}{\sqrt{p_{\bar{D}k}}}\left\{R_1^k(Y_{1\bar{D}i}^k)-R_2^k(Y_{2\bar{D}i}^k)-E(R_1^k-R_2^k)\right\}\\
&+\frac{1}{\sqrt{\lambda n_D}}\sum_{j=1}^{n_D}\left\{Q_{1Xj}-Q_{2Xj}-E(Q_{1X}-Q_{2X})\right\}+o_p(1).
\end{aligned}
$$

This proves Theorem 4(a). Now consider when F is estimated parametrically. Let $X_{\bar{D}i}, i=1,...,n_{\bar{D}}$ and $X_{Dj}, j=1,...,n_D$ be the covariate measurements for controls and cases. Let $F_{\theta_z}(.|X)$ be the distribution function of marker z in controls conditional on X. Suppose $\sqrt{n}\left(\hat{\theta}-\theta\right)$ can be represented as $\frac{1}{\sqrt{n_{\bar{D}}}}\sum_{i=1}^{n_{\bar{D}}}\psi_i+o_p(1)$, where $\psi_i, i=1,...,n_{\bar{D}}$ are independent identically distributed variables with $E(\psi_i)=0$ and var$(\psi_i)=\Sigma(\theta)$, we have

$$
\begin{aligned}
&\sqrt{n_{\bar{D}}}\left(\hat{\Delta}-\Delta\right)\\
&=\sqrt{n_{\bar{D}}}\left[\frac{1}{n_D}\sum_{j=1}^{n_D}\left\{F_{\hat{\theta}_1}(Y_{1D_j}|X_{Dj})-F_{\hat{\theta}_2}(Y_{2D_j}|X_{Dj})\right\}-E_{Y_{1D},X_D}F_{\theta_1}(Y|X)+E_{Y_{2D},X_D}F_{\theta_2}(Y|X)\right]\times 100\\
&=\sqrt{n_{\bar{D}}}\left\{E_{Y_{1D},X_D}F_{\hat{\theta}_1}(Y_{1D}|X_D)-E_{Y_{1D},X_D}F_{1,\theta}(Y_{1D}|X_D)-E_{Y_{2D},X_D}F_{\hat{\theta}_2}(Y_{2D}|X_D)\right.\\
&\left.+\ E_{Y_{2D},X_D}F_{\theta_2}(Y_{2D}|X_D)\right\}\times 100+\sqrt{n_{\bar{D}}}\left[\frac{1}{n_D}\sum_{j=1}^{n_D}\left\{F_{\theta_1}(Y_{1D_j}|X_{Dj})-F_{\theta_2}(Y_{2D_j}|X_{Dj})\right\}\right.\\
&\left.-\ E_{Y_{1D},X_D}F_{\theta_1}(Y|X)+E_{Y_{2D},X_D}F_{\theta_2}(Y|X)\right]\times 100+o_p(1)\\
&=\frac{1}{\sqrt{n_{\bar{D}}}}\sum_{i=1}^{n_{\bar{D}}}\frac{\partial\Delta_i}{\partial\theta}\psi_i+\frac{1}{\sqrt{\lambda n_D}}\sum_{j=1}^{n_D}\left\{Q_{1Xj}-Q_{2Xj}-E(Q_{1X}-Q_{2X})\right\}+o_p(1).
\end{aligned}
$$

0.4.*1* *Theorem 4 (Extension Version)*

Consider the setting when the two markers are adjusted for different covariates X_1 and X_2. For $z=1,2$, let $Q_{zX_z}(\hat{Q}_{zX_z})$ be the (estimated) covariate-specific percentile value for the z^{th} marker, let $\Delta=E(Q_{1X_1})-E(Q_{2X_2})$ and $\hat{\Delta}=\overline{\hat{Q}_{1X_1}}-\overline{\hat{Q}_{2X_2}}$. When X_1 and X_2 are discrete with strata $k_1=1,\ldots,K_1$ and $k_2=1,\ldots,K_2$ respectively, let $n_{Dk_1k_2}$ and $n_{\bar{D}k_1k_2}$ be the number of controls and cases in the intersection of the k_1^{th} stratum for covariate 1 and the k_2^{th} stratum for covariate 2. Suppose as $n_{\bar{D}}\to\infty, n_D/n_{\bar{D}}\to\lambda\in(0,1)$, $n_{\bar{D}k_1k_2}/n_{\bar{D}}\to p_{\bar{D}k_1k_2}\in(0,1)$, $n_{Dk_1k_2}/n_D\to p_{Dk_1k_2}\in(0,1)$, then

$\sqrt{n_{\bar{D}}}\left(\hat{\Delta}-\Delta\right)$ converges to a mean 0 normal random variable with variance σ^2, where

$$(a) \qquad \sigma^2=\sum_{k_1}\sum_{k_2}\frac{\text{var}\left\{R_1^{k_1}(Y_{1\bar{D}}^{k_1})-R_2^{k_2}(Y_{2\bar{D}}^{k_2})\right\}}{p_{\bar{D}k_1k_2}/p_{\bar{D}k_1k_2}^2}+\frac{\text{var}(Q_{1X_1}-Q_{2X_2})}{\lambda},$$

if for each marker, the covariate-specific reference distribution $F(Y|X)$ is estimated empirically within each covariate category specific for the marker, where $R_z^{k_z}\left(Y_{z\bar{D}}^{k}\right)=P(Y_{zD}^{k_z}<Y_{z\bar{D}}^{k_z})$ is the percentile value for a control using his covariate-specific case distribution as the reference for the z^{th} marker in the k_z^{th} covariate category, and

$$(b) \qquad \sigma^2=\left(\frac{\partial\Delta}{\partial\theta}\right)^T\Sigma(\theta)\left(\frac{\partial\Delta}{\partial\theta}\right)+\frac{\text{var}(Q_{1X_1}-Q_{2X_2})}{\lambda}, \tag{0.1}$$

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.

Proof follows similar steps to above.

0.5 *Proof of Proposition 1*

Let $Y_{\bar{D}i}, i=1,...,n_{\bar{D}}$, and Y_z be the marker measurement in controls and the z^{th} type of cases, $z=1,2$. Let $Q_z(\hat{Q}_z)$ be the corresponding (estimated) case percentile value. We have

$$\begin{aligned}
Q_1\overset{d}{=}Q_2 &\Rightarrow P\{F(Y_1)\leqslant t\}=P\{F(Y_2)\leqslant t\} \qquad \forall t\in(0,1)\\
&\Rightarrow P\{Y_1\leqslant F^{-1}(t)\}=P\{Y_2\leqslant F^{-1}(t)\}\\
&\Rightarrow Y_1\overset{d}{=}Y_2\\
&\Rightarrow \hat{Q}_1|Y_{\bar{D}i}\overset{d}{=}\hat{Q}_2|Y_{\bar{D}i}, i=1,...,n_{\bar{D}} \qquad \text{and} \qquad \hat{Q}_1\overset{d}{=}\hat{Q}_2.
\end{aligned}$$

0.6 *Proof of Proposition 2*

Let $Y_{\bar{D}}$ and Y_D be the marker measurements for controls and cases respectively, $Y_{\bar{D}}\sim F, Y_D\sim G$. Let $\hat{Q}$ be the estimated case percentile value. When $\hat{F}$ is the empirical CDF, for $m\in(0,1,...,n_{\bar{D}})$,

$$\begin{aligned}
&P(n_{\bar{D}}\hat{Q}/100=m)=P\left\{n_{\bar{D}}\hat{F}(Y_D)=m\right\}=P\left\{\sum_{i=1}^{n_{\bar{D}}}I(Y_{\bar{D}}\leqslant Y_D)=m\right\}\\
&=\frac{n_{\bar{D}}!}{(n_{\bar{D}}-m)!m!}\int_{-\infty}^{\infty}F^m(y)\{1-F(y)\}^{n_{\bar{D}}-m}dG(y)\\
&=\frac{n_{\bar{D}}!}{(n_{\bar{D}}-m)!m!}\int_0^1 w^m(1-w)^{n_{\bar{D}}-m}dG\{F^{-1}(w)\},
\end{aligned}$$

which is a function of ROC. Therefore, when $\text{ROC}_1(t)=\text{ROC}_2(t)\ \forall t$, i.e. when $Q_1\overset{d}{=}Q_2$, $\hat{Q}_1$ and $\hat{Q}_2$ have the same distribution.

0.7 *Proof of Proposition 3*

With $z = 1, 2$, let $\hat{Q}_{zi}, i = 1, ..., n_D$ be the estimated case percentile values for the z^{th} marker. Because of the exchangeability among cases, the correlation between $\left(\hat{Q}_{1i}, \hat{Q}_{2i}\right)$ and $\left(\hat{Q}_{1j}, \hat{Q}_{2j}\right)$ are the same for each pair of i, j. In another words, $\left(\hat{Q}_{11}, \hat{Q}_{21}\right), \ldots, \left(\hat{Q}_{1n_D}, \hat{Q}_{2n_D}\right)$ are correlated samples from a bivariate distribution with exchangeable correlation. Under $H_0 : Q_1 \stackrel{d}{=} Q_2$, $\hat{Q}_1$ and $\hat{Q}_2$ have the same distribution when F is estimated empirically (Proposition 2). Consequently $U_j = \hat{Q}_{1j} - \hat{Q}_{2j}, j = 1, ..., n_D$ are correlated (exchangeably) samples from a symmetric univariate distribution H. Due to the complete exchangeability among those $U_j's$, if H is continuous, $Z_j = P(U_j > 0)$ is Bernoulli random variable with $P(Z_j = 1) = 0.5$, thus $E(S) = \sum_{j=1}^{n_D} Z_j$ is equal to $1/2$. In addition, Z_j is independent of $r(|U_j|)$, where $r(|U_j|)$ is the rank of $|U_j|$ in the sample, thus, $E(T) = \sum_{j=1}^{n_D} E\left\{Z_j r(|U_j|)\right\} = \sum_{j=1}^{n_D} \frac{1}{2} E\{r(|U_j|)\} = (n_D + 1)/4$.

0.8 *Proof of Proposition 4*

It is equivalent to prove that the Mann-Whitney test statistic (MW) of the two groups of estimated percentile values has mean 0.5 when F is estimated empirically. We have

$$MW = \frac{1}{n_D^2} \sum_{i=1}^{n_D} \sum_{j=1}^{n_D} \left\{ I\left(\hat{Q}_{1i} < \hat{Q}_{2j}\right) + \frac{I\left(\hat{Q}_{1i} = \hat{Q}_{2j}\right)}{2} \right\}$$

$$E(MW) = \frac{1}{n_D^2} \sum_{i=1}^{n_D} \sum_{j=1}^{n_D} \left\{ P\left(\hat{Q}_{1i} < \hat{Q}_{2j}\right) + \frac{1}{2} P\left(\hat{Q}_{1i} = \hat{Q}_{2j}\right) \right\}$$

$$= P\left(\hat{Q}_{1i} < \hat{Q}_{2j}\right) + \frac{P\left(\hat{Q}_{1i} = \hat{Q}_{2j}\right)}{2}.$$

As proved in Proposition 2, when $\text{ROC}_1(t) = \text{ROC}_2(t)\ \forall t$, i.e. when $Q_1 \stackrel{d}{=} Q_2$, $\hat{Q}_1$ and $\hat{Q}_2$ have the same distribution when $\hat{F}$ is estimated empirically. Therefore, $E(MW) = 0.5$.

REFERENCES

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**(3), 837-845.

VAN DER VAART, A. W. (1998) *Asymptotic Statistics*. Cambridge University Press.

VAN DER VAART, A. W. AND WELLNER, J. A. (1996) *Weak Convergence and Empirical processes*. Springer-Verlag, New York.